HANDBOOK OF MATHEMATICAL STATISTICS

BY

H. L. RIETZ, Editor-in-Chief

H. C. CARVER
A. R. CRATHORNE
W. L. CRUM
JAMES W. GLOVER
E. V. HUNTINGTON
TRUMAN L. KELLEY
WARREN M. PERSONS
ALLYN A. YOUNG

Members of the Committee on the Mathematical Analysis of Statistics of the Division of Physical Sciences of the National Research Council

HOUGHTON MIFFLIN COMPANY
BOSTON · NEW YORK · CHICAGO · DALLAS · SAN FRANCISCO
The Riverside Press Cambridge

The Riverſide Preſs
CAMBRIDGE · MASSACHUSETTS
PRINTED IN THE U.S.A.

PREFACE

The study of a problem by statistical methods usually involves three stages: (1) the collection of material or data; (2) the mathematical analysis of the data thus collected; (3) the interpretation of results, for the particular purpose in view.

As to stage (1), the best methods of collecting data depend almost entirely on the nature of the particular field of inquiry, and are not discussed in this Handbook. The same is true in regard to stage (3); the problems connected with the interpretation of statistical results are necessarily very different in different fields of inquiry, and are not discussed in this Handbook, except as illustrations of the mathematical methods involved.

The problems of stage (2), on the other hand, are in a sense common to all fields of statistical inquiry. Whatever the content of the data may be, the *form* of the mathematical analysis is essentially the same. It is with these formal problems of mathematical analysis that this Handbook deals. Illustrations are taken from this or that particular field, for the sake of concreteness; but the *general* applicability of the methods to all fields is constantly borne in mind, and the terminology throughout the Handbook is kept as non-special as possible.

Special emphasis is laid on the limitations surrounding the proper application of the various methods of analysis. Without careful attention to these limitations, the results of a statistical inquiry may be altogether misleading.

Each chapter has been critically read by at least two other contributors besides the author; but the final responsibility for all the chapters rests with the individual authors.

The National Research Council contributed to the preparation of the Handbook by the grant of funds for traveling expenses incident to meetings of the Committee and for a small amount of clerical assistance. The royalties from the book are received by the National Research Council to be made available, if needed, for further work in the field of mathematical statistics.

CONTENTS

CHAPTER I. MATHEMATICAL MEMORANDA

By E. V. Huntington

Professor of Mechanics, Harvard University

CHAPTER II. FREQUENCY DISTRIBUTIONS, AVERAGES, AND MEASURES OF DISPERSION (ELEMENTARY METHODS)

By H. L. Rietz

Professor of Mathematics, University of Iowa

CHAPTER III. INTERPOLATION, SUMMATION, AND GRADUATION

By James W. Glover

Professor of Mathematics and Insurance, University of Michigan

CHAPTER IV. CURVE-FITTING BY THE METHOD OF LEAST SQUARES AND THE METHOD OF MOMENTS

By E. V. Huntington

CHAPTER V. RANDOM SAMPLING

By H. L. Rietz

CHAPTER VI. BERNOULLI, POISSON, AND LEXIS DISTRIBUTIONS

By H. L. Rietz

CHAPTER VII. FREQUENCY CURVES

By H. C. Carver

Professor of Mathematics and Insurance, University of Michigan

CHAPTER VIII. SIMPLE CORRELATION

By H. L. Rietz
and
A. R. Crathorne

Associate Professor of Mathematics, University of Illinois

CHAPTER IX. PARTIAL AND MULTIPLE CORRELATION

By Truman L. Kelley

Professor of Education, Stanford University

CHAPTER X. CORRELATION OF TIME SERIES

By Warren M. Persons

Professor of Economics, Harvard University

CHAPTER XI. PERIODOGRAM ANALYSIS

By W. L. Crum

Assistant Professor of Statistics, Stanford University

CHAPTER XII. INDEX NUMBERS

By Allyn A. Young

Late Professor of Economics, Harvard University

HANDBOOK OF MATHEMATICAL STATISTICS

CHAPTER I

MATHEMATICAL MEMORANDA

By E. V. HUNTINGTON

NUMERICAL COMPUTATION

Slide rules, tables, and computing machines. Before undertaking any statistical work one should supply one's self with suitable aids to computation.

For three-figure accuracy, a ten-inch **slide rule** is very convenient. The larger Fuller or Thacher slide rules give four, or sometimes five, significant figures. Barlow's Tables of squares, square roots, cube roots, and reciprocals, are almost indispensable. The multiplication tables are also often convenient. Crelle's Table gives the product of every three-figure number by every three-figure number. Peters's Table gives the product of every four-figure number by every two-figure number. The smaller table of H. Zimmermann gives the product of every three-figure number by every two-figure number. Tables of logarithms of numbers, and for certain purposes tables of trigonometric functions, are invaluable. Four- and five-place tables exist in great variety. If more than five figures are required, use Bremiker's six-place table or proceed at once to a seven-place table: for example, Vega. For eight places use the two-volume table of Bauschinger and Peters. Explanations of the use of tables of logarithms usually accompany the tables themselves; see, for example, E. V. Huntington's *Handbook of Mathematics for Engineers.*

In extended work some form of computing machine will soon pay for itself in spite of the apparently large initial expense. The best-known adding and listing machines are the Burroughs and the Wales, with standard keyboards, and the Dalton and the Sundstrand with ten-key keyboards. (The wide-paper form of carriage is more convenient for most purposes than the narrow-ribbon type.) Among the calculating machines may be mentioned the Comptometer, the Burroughs nonlisting machine, the Monroe calculator, the Millionaire, the Brunsviga, the Ensign, the Mercedes-Euclid, and the Marchant. Some of these can be operated by electricity.

For elaborate classification of large amounts of statistical data, as in the work of the Census Bureau, the Hollerith or the Powers machine for sorting punched cards is practically indispensable.

In advanced work in statistical theory, Pearson's *Tables for Statisticians and Biometricians* are invaluable.

The new *Tables for Applied Mathematics*, by J. W. Glover, include in one volume a large number of tables for finance, insurance, and statistics, together with a seven-place table of logarithms.

Absolute and relative errors. The numerical data in a statistical computation are usually the result of measurement, observation, or estimate, and hence are only approximately correct. The closeness of the approximation may be measured either by the absolute error or by the relative error.

The **absolute error** is sometimes defined as the observed value minus the true value ($x_1 - X$) and sometimes as the true value minus the observed value ($X - x_1$). When the distinction of sign is important, the error $x_1 - X$ may be called the **deviation** of the observed value from the true value (a positive deviation being an " error in excess," and a negative deviation an " error in defect "), while the error $X - x_1$ may be called the **correction** to be applied to the observed quantity (the correction being positive or negative according as the observed quantity needs to be increased or decreased).

The **relative error** is the absolute error divided by the true value.

For example, suppose $x_1 = 3.06$ cm. and $x_2 = 2.97$ cm. are two approximate values and $X = 3.00$ cm. is the true value. Then the absolute error of x_1 is 0.06 cm. (deviation = + 0.06 cm., correction = − 0.06 cm.) while the relative error is 0.02, or 2 per cent. Similarly, the absolute error of x_2 is 0.03 cm. (deviation = − 0.03 cm., correction = + 0.03 cm.) while the relative error is 0.01, or 1 per cent.

The absolute error is connected with the **number of decimal places,** and is important when the quantity is to be added or subtracted, or compared with other quantities on an absolute basis. For example, a measurement may be " correct to two decimal places "; an estimated population may be " correct to the nearest million," etc.

The relative error, on the other hand, is connected with the **number of significant figures,** and is important when the quantity is to be multiplied or divided, or compared with another quantity on a percentage basis. For example, a number may be said to be " correct to four significant figures," "correct to within 3 per cent of the value," "correct within one part in 6000," etc.

In any statistical investigation, either the desired number of decimal places, or more usually, the desired number of significant figures should be decided upon in advance, and borne constantly in mind throughout the work.

Propagation of error in computation. The manner in which small errors in the data may accumulate in the course of a computation is indicated as follows:

(1) In addition: Suppose, for example, that each of 20 numbers has a possible error of half a unit in the third decimal place; then the sum of these numbers may have a possible error of 10 units in the third decimal place — that is, an error of 1 unit in the second decimal place. All figures beyond the second decimal place should therefore be discarded in the answer. In general, one doubtful figure in any column will render that whole column doubtful; hence all figures to the right of that column should be discarded in the answer.

(2) In subtraction: Two numbers may each be correct, say, to five significant figures, and yet their difference may be correct to only one or two significant figures; for example, $3.1416 - 3.1402 = 0.0014$. Neglect of this fact is a frequent source of overconfidence in regard to the precision of a result.

(3) In multiplication and division: The number of significant figures which can be relied on in a product or quotient is never greater than the number of reliable significant figures in the weakest factor.

The relative error of a product or quotient may be as great as the sum of the relative errors of the separate items.

(4) In powers and roots: The relative error in the nth power of a number is n times the relative error in the number itself. Similarly, the relative error in $\sqrt[n]{x}$ is only $1/n$th of the relative error in x itself.

(5) In exponents and logarithms: If $y = e^x$, or $x = \log y$, then an absolute error of say .01 in x corresponds approximately to a relative error of .01 in y.

(6) In arithmetic or geometric mean: The relative error in the arithmetic or geometric mean of a number of quantities will be approximately the same as the relative error of the individual items (greater than the least of these relative errors and less than the greatest of them).

Rejection of superfluous figures. It is a fundamental rule of computation that a result should never be stated to a greater degree of precision than is justified by the data. All superfluous digits are misleading and should be rejected from the result.

If the first rejected figure is 5 or more, the preceding figure should be increased by one; otherwise, it should be left unchanged.[1]

For example, 3.14159 reduced to four figures is 3.142. Again, 6.1297 reduced to four figures is 6.130. Note that in a decimal fraction a final zero is as significant as any other final digit in determining the degree of precision. But in the case of a whole number like 3140000 the final zeros leave the reader in doubt whether the number of reliable significant figures is 3, 4, 5, 6, or 7. This ambiguity can be removed by writing the number in the form 3140000, or 3140000, or 3140000, etc., as the case may require; or, more usually, in the form 3.14×10^6, or 3.140×10^6, or 3.1400×10^6, etc., as the case may require.

This latter "**notation by powers of 10**" should always be used in the case of very large or very small numbers. For example,

$$0.000003140 = 3.140 \times 10^{-6}.$$

(*Note:* In this notation, the exponent of the power of 10 is the same as the "characteristic" of the logarithm of the number.)

DEFINITIONS OF VARIOUS KINDS OF MEANS OR AVERAGES

(1) The **arithmetic mean** (AM) of n numbers, $x_1, x_2, \cdots x_n$, is $1/n$th of their sum:

$$AM = \frac{1}{n}(x_1 + x_2 + \cdots + x_n), \text{ or } AM = \frac{1}{n}\Sigma x_i.$$

The AM is what is ordinarily meant when the term "mean" or "average" is used without further qualification. It is related to the **center of gravity** (or **centroid**) in mechanics, the center of gravity of a set of n equal particles being a point whose distance from any fixed plane is the AM of the distances of the several particles from that plane. It is also related to the "**method of least squares,**" since the sum of the squares of the deviations of the n numbers from any value X is a minimum when X is the AM of the numbers.

In computing an AM note that adding any constant, $\pm h$, to all the numbers has the effect of adding $\pm h$ to their AM.

For two numbers, a and b, the $AM = \frac{1}{2}(a + b)$.

(2) The **geometric mean** (GM) of n (positive) numbers, $x_1, x_2, \cdots x_n$, is the nth root of their product:

$$GM = \sqrt[n]{x_1,x_2, \cdots x_n}.$$

[1] A refinement of this rule is sometimes to be recommended, namely: if the rejected figures are exactly 5000 $\cdots$, the preceding figure should be raised when it is odd and left unchanged when it is even.

In computing the GM of n numbers, it is usually convenient to use the formula:

$$\log(GM) = \frac{1}{n}(\log x_1 + \log x_2 + \cdots + \log x_n), \text{ or } \log(GM) = \frac{1}{n}\Sigma(\log x_i),$$

that is, take the AM of the logarithms of the numbers, and then take the anti-log of the result.

For two numbers, a and b, the GM is $x = \sqrt{ab}$. This is called also the mean proportional between a and b, since $a : x = x : b$. By drawing a semicircle on $a + b$ as diameter, the value of x can be constructed geometrically, as in Figure 1.

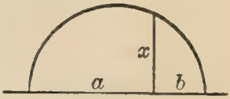


FIG. 1

(3) The **harmonic mean** (HM) of n (positive) numbers, $x_1, x_2, \cdots x_n$, is the reciprocal of the arithmetic mean of the reciprocals of the numbers:

$$HM = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_3}\right)}, \text{ or } \frac{1}{HM} = \frac{1}{n}\Sigma\left(\frac{1}{x_i}\right).$$

The chief use of the HM is in averaging frequencies, $1/x$ being called a frequency when x is a duration.

For example, in steamship statistics, the average number of trips per year may be more significant than the average number of days spent on each trip.

For two numbers, a and b, $HM = \dfrac{2\,ab}{a + b}$.

The AM, GM, and HM are the so-called classical means, known to the Greeks.

(4) The **contra-harmonic mean** (CHM) is almost as old, but is of very slight importance to-day:

$$CHM = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{x_1 + x_2 + \cdots + x_n} \text{ or } CHM = \Sigma(x_i^2)/\Sigma(x_i).$$

(5) The **root-mean-square** (RMS) of n numbers, $x_1, x_2, \cdots x_n$, is the square root of the arithmetic mean of their squares:

$$RMS = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2)}, \text{ or } RMS = \sqrt{\frac{1}{n}\Sigma(x_i^2)}.$$

The RMS is related to the **radius of gyration** in mechanics, the radius of gyration of a set of n equal particles, with respect to a given axis, being the RMS of the radial distances of the several particles from that axis. In statistics, the RMS of the deviations of a set of numbers from their arithmetic mean is called the **standard deviation** (SD) of those numbers.

The standard deviation of a set of numbers is also equal to the *RMS* of the differences between the numbers taken two by two; thus, if x is the *AM* of the numbers, then

$$SD = \sqrt{\frac{1}{n}\sum_{i}(x_i - x)^2} = \frac{1}{n}\sqrt{\sum_{i,j}(x_i - x_j)^2},$$

where $n(n-1)/2 =$ the number of the differences in question.

For any positive numbers, $x_1 \leqq x_2 \leqq \cdots \leqq x_n$, the order of magnitude of these five means is as follows (unless the numbers are all equal):

$$x_1 < HM < GM < AM < RMS < CHM < x_n.$$

For the special case of two numbers, a and b, the following facts may be noted:

The *GM* of two numbers is the *GM* between their *HM* and their *AM*.

The *AM* of two numbers is the *AM* between their *HM* and their *CHM*.

The *RMS* of two numbers is the *GM* between their *AM* and their *CHM*.

The following general formulas, due to Mr. R. M. Foster, may also be noted:

$$M = [(x_1{}^k + x_2{}^k + \cdots + x_n{}^k)/n]^{1/k}$$

$$M' = \frac{x_1{}^{k+1} + x_2{}^{k+1} + \cdots + x_n{}^{k+1}}{x_1{}^k + x_2{}^k + \cdots + x_n{}^k}$$

If $k =$	$-\infty$	-1	0	1	2	∞
then $M =$	x_1	*HM*	*GM*	*AM*	*RMS*	x_n
and $M' =$	x_1	*HM*	*AM*	*CHM*		x_n

(The proof involves the evaluation of certain simple indeterminate forms.)

(6) The **median** of a set of quantities is, roughly speaking, the middle one of the set, when they are arranged in order of magnitude (*i.e.* "arrayed"). If the number of quantities is even, and the two middle quantities are not equal, the median is commonly taken as the number halfway between them. More exactly, the median, in this case, is a number X uniquely determined by the equation

$$(X - a_1)(X - a_2)\cdots(X - a_k) = (a_{k+1} - X)(a_{k+2} - X)\cdots(a_n - X),$$

where $a_1, a_2, \cdots a_k$ are the quantities of the lower half, and $a_{k+1}, a_{k+2}, \cdots a_n$, the quantities of the upper half of the set. (D. Jackson, *Bull. Amer. Math. Soc.*, Jan. 1921.)

The sum of the absolute deviations of n numbers from any value X is a minimum when X is the median of those numbers.

(7) The **mode** of a set of quantities is that quantity which occurs most often (*i.e.* is the most fashionable), if such a quantity exists. Any quantity which occurs more often than any other quantity near it in size may be called a relative mode (or simply *a* mode) of the set.

A general mathematical formula including the arithmetic mean, the median, and the mode is due to D. Jackson and R. M. Foster: Let X be the value of x which minimizes $\Sigma|x_i - x|^p$. Then if $p = 2$, $X =$ the arithmetic mean; if $p \to 1$, $\lim X =$ the median; if $p \to 0$, $\lim X =$ the mode. We note also that if $p \to \infty$, $\lim X = \frac{1}{2}(x_1 + x_n)$, where x_1 is the smallest and x_n the largest of the given quantities.

In the case of the median and the mode (as in the case of the AM), adding a constant, $\pm h$, to all the numbers has the effect of adding the same constant to the mean. (This is not true in case of the other four types of means.)

The following general properties are often useful:

In the case of any one of the seven means, multiplying all the numbers by a constant factor, c, has the effect of multiplying the mean by the same constant, c. ("Change of scale.")

In computing the AM, GM, HM, or RMS of n numbers, it is allowable, after grouping the numbers in any way, to replace each number of any group by the corresponding mean of that group. (This is not allowable in the case of the CHM, the median, or the mode.)

Weighted means. If the given numbers $x_1, x_2, \cdots x_n$ have different degrees of importance, as indicated by "weights" $w_1, w_2, \cdots w_n$, then we may speak of the **weighted mean** of these numbers (of any one of the seven kinds). Any kind of weighted mean of the given set of n numbers is defined as the corresponding kind of simple mean of a set of W numbers, in which x_1 occurs w_1 times, x_2 occurs w_2 times, etc., and $W = w_1 + w_2 + \cdots + w_n$ is the sum of the weights.

For example, the weighted arithmetic mean is $\frac{1}{W}(w_1x_1 + w_2x_2 + \cdots + w_nx_n)$; the weighted geometric mean is $(x_1^{w_1}x_2^{w_2} \cdots x_n^{w_n})^{1/W}$; etc.

The term "weighted mean," or "weighted average," used without qualifying adjective, usually indicates the weighted arithmetic mean.

PERMUTATIONS AND COMBINATIONS. THE BINOMIAL THEOREM

Permutations. The number of possible permutations or arrangements of n different elements is "n factorial" $= n! = 1 \cdot 2 \cdot 3 \cdots n$. Another notation is $\underline{|n} = n!$

Thus, the three letters a, b, c admit $3! = 6$ permutations: *abc, acb, bac, bca, cab, cba.*

If among the n elements there are p equal ones of one sort, q equal ones of another sort, r equal ones of a third sort, etc., where $p + q + r + \cdots = n$, then the number of possible permutations is

$$(n!) \,/\, [(p!)(q!)(r!)\cdots].$$

Thus, the four letters a, b, b, b, admit $24/[(1)(6)] = 4$ permutations: *abbb, babb, bbab, bbba.*

Combinations. The number of possible combinations or groups of n elements taken r at a time (without repetition of any element within any one group) is ${}_nC_r = \dfrac{n!}{(n-r)!\,r!}$ = the coefficient of the term in x^r in the binomial expansion of $(1 + x)^n$. (Notice that ${}_nC_r = {}_nC_{n-r}$).

Thus, the five letters *abcde* taken two at a time give ${}_5C_2 = 10$ combinations: *ab, ac, ad, ae, bc, bd, be, cd, ce, de.*

If repetitions are allowed within each group, then the number of combinations of n elements taken r at a time is ${}_{n+r-1}C_r$.

Thus, five letters taken two at a time, repetitions allowed, give ${}_6C_2 = 15$ combinations: *aa, ab, ac, ad, ae, bb, bc, bd, be, cc, cd, ce, dd, de, ee.*

The general principle underlying the theory of permutations and combinations is this: If we can do one thing in m ways and another thing in n ways, then we can do both things together in mn ways.

The binomial theorem. If n is any positive integer,

$$(p+q)^n = p^n + np^{n-1}q + \frac{n(n-1)}{1\cdot 2}p^{n-2}q^2 + \frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}p^{n-3}q^3 + \cdots + q^n$$

$$= p^n + {}_nC_1p^{n-1}q + {}_nC_2p^{n-2}q^2 + {}_nC_3p^{n-3}q^3 + \cdots + q^n,$$

where ${}_nC_1 = n$, ${}_nC_2 = [n(n-1)]/(2!)$, ${}_nC_3 = [n(n-1)(n-2)]/(3!)$, $\cdots$

$${}_nC_r = [n(n-1)(n-2)\cdots(n-r+1)]/(r!)$$

are the **binomial coefficients.**

Note that ${}_nC_{n-r} = {}_nC_r = \dfrac{n!}{(n-r)!\,r!}$.

Other notations are ${}_nC_r = \dbinom{n}{r} = (n)_r$.

Table of Binomial Coefficients

n	${}_nC_0$	${}_nC_1$	${}_nC_2$	${}_nC_3$	${}_nC_4$	${}_nC_5$	${}_nC_6$	${}_nC_7$	${}_nC_8$	${}_nC_9$	${}_nC_{10}$
1	1	1	..	..	..	..	..	..	..	..	..
2	1	2	1	..	..	..	..	..	..	..	..
3	1	3	3	1	..	..	..	..	..	..	..
4	1	4	6	4	1	..	..	..	..	..	..
5	1	5	10	10	5	1	..	..	..	..	..
6	1	6	15	20	15	6	1	..	..	..	..
7	1	7	21	35	35	21	7	1	..	..	..
8	1	8	28	56	70	56	28	8	1	..	..
9	1	9	36	84	126	126	84	36	9	1	..
10	1	10	45	120	210	252	210	120	45	10	1
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

Note that each number, plus the number on its left, gives the number next below.

STIRLING'S FORMULA. THE BERNOULLI NUMBERS

Stirling's formula for n factorial. The following formula gives a good approximation to $n!$ for large values of n:

$$n! = (\sqrt{2\pi n})(n^n)(e^{-n}), \text{ or, more accurately,}$$

$$n! = (\sqrt{2\pi n})(n^n)(e^{-n})(e^{\frac{\theta}{12n}}), \text{ where } 0 < \theta < 1,$$

whence

$$\log_e (n!) = (n + \tfrac{1}{2}) \log_e n - n + \log_e (\sqrt{2\pi}) + \frac{\theta}{12\,n},$$

or $\log_{10} (n!) = (n + \frac{1}{2}) \log_{10} n - (.434294482\,n) + .39909 + \dfrac{.03619\,\theta}{n}.$

The last term, in which $0 < \theta < 1$, indicates the degree of approximation attained. For example, if $n = 1000$, $\log_{10} (1000!) = 2567.6046$, so that $1000! = 4.024 \times 10^{2567}$.

A seven-place table of $\log_{10} (n!)$ up to $n = 1000$ is given in *Pearson's Tables*, page 98, and in *Glover's Tables*, page 482.

A still more accurate approximation is

$$\log_e (n!) = (n + \tfrac{1}{2}) \log_e n - n + \log_e (\sqrt{2\pi}) + \frac{B_1}{1 \cdot 2}\frac{1}{n} - \frac{B_3}{3 \cdot 4}\frac{1}{n^3} + \frac{B_5}{5 \cdot 6}\frac{1}{n^5} - \frac{B_7}{7 \cdot 8}\frac{1}{n^7} + \frac{\theta B_9}{9 \cdot 10}\frac{1}{n^9},$$

where $B_1 = \frac{1}{6}$, $B_3 = \frac{1}{30}$, $B_5 = \frac{1}{42}$, $\cdots$ are the Bernoulli numbers (see below), and $0 < \theta < 1$.

Wallis's formula for π. Wallis's formula (useful in connection with the proof of Stirling's formula) is an infinite product the limit of which is $\pi/2$:

$$\frac{\pi}{2}=\frac{2}{1}\,\frac{2}{3}\,\frac{4}{3}\,\frac{4}{5}\,\frac{6}{5}\,\frac{6}{7}\cdots\frac{2\,n}{2\,n-1}\,\frac{2\,n}{2\,n+1}\cdots$$

The Bernoulli numbers. The following numbers occur in the expansion of many functions, such as tan x, sec x, $x/(e^x-1)$, etc.

$B_1 = 1/6$	$B_2 = 1$
$B_3 = 1/30$	$B_4 = 5$
$B_5 = 1/42$	$B_6 = 61$
$B_7 = 1/30$	$B_8 = 1385$
$B_9 = 5/66$	$B_{10} = 50521$
$B_{11} = 691/2730$	$B_{12} = 2702765$
$B_{13} = 7/6$	$B_{14} = 199360981$
$B_{15} = 3617/510$	$B_{16} = 19391512145$
$B_{17} = 43867/798$	$B_{18} = 2404879675441$
$B_{19} = 174611/330$	$B_{20} = 370371188237525$
etc.	etc.

The numbers $B_2, B_4, B_6, \cdots$ are sometimes denoted by $E_2, E_4, E_6, \cdots$ or by $E_1, E_2, E_3, \cdots$; while the numbers $B_1, B_3, B_5, \cdots$ are sometimes denoted by $B_2, B_4, B_6, \cdots$ or by $B_1, B_2, B_3, \cdots$

For recursion formulas, see B. O. Peirce, *Table of Integrals.* For an extended table, see *Glover's Tables.* For large values of n, the following approximations are useful:

$$\frac{B_{2n-1}}{(2\,n)\,!}=\frac{2}{(2^{2n}-1)\pi^{2n}}\left[1+\frac{1}{3^{2n}}+\frac{1}{5^{2n}}+\frac{1}{7^{2n}}+\cdots\right]$$

$$\frac{B_{2n}}{(2\,n)\,!}=\frac{2^{2n+2}}{\pi^{2n+1}}\left[1-\frac{1}{3^{2n+1}}+\frac{1}{5^{2n+1}}-\frac{1}{7^{2n+1}}+\cdots\right]$$

THE GAMMA FUNCTION

The **Gamma Function** of any positive number n is defined by

$$\Gamma(n)=\int_0^\infty x^{n-1}e^{-x}dx.$$

If n is a positive integer, $\Gamma(n+1)=n\,!$. (See Stirling's formula, above.) In general, $\Gamma(n+1)=n\,\Gamma(n)$, so that the value of $\Gamma(n)$ for any positive n can be found, by successive reductions, from a table covering the range from any integer to the succeeding integer, as, for example, from $n=1$ to $n=2$. In particular,

$$\Gamma(0)=\infty,\quad \Gamma(\tfrac{1}{2})=\sqrt{\pi},\quad \Gamma(1)=1,\quad \Gamma(2)=1,\quad \Gamma(3)=2.$$

The graph of the function is shown in Figure 2. The minimum point is given by $\Gamma(1.4616321) = .8856032$. Tables of the Gamma Function are given in *Pearson's Tables* and in *Glover's Tables.*

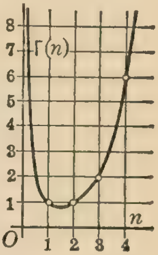


FIG. 2

The Beta Function. The Beta Function of any two positive numbers, m and n, is defined by

$$B(m,n) = \int_0^1 x^{m-1}(1-x)^{n-1}dx$$

$$= 2\int_0^{\frac{\pi}{2}} \sin^{2m-1}\theta \cdot \cos^{2n-1}\theta \cdot d\theta = \frac{\Gamma(m)\,\Gamma(n)}{\Gamma(m+n)}.$$

The hypergeometric series. The hypergeometric series is a function of x involving three parameters, a, b, c:

$$F(a,b,c,x) = 1 + \frac{a\cdot b}{1\cdot c}x + \frac{a(a+1)}{1\cdot 2}\,\frac{b(b+1)}{c(c+1)}x^2$$

$$+ \frac{a(a+1)(a+2)}{1\cdot 2\cdot 3}\,\frac{b(b+1)(b+2)}{c(c+1)(c+2)}x^3 + \cdots$$

$$= \frac{\Gamma(c)}{\Gamma(a)\cdot\Gamma(c-a)}\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-xt)^{-a}dt.$$

GAUSS'S NORMAL ERROR CURVE, OR PROBABILITY CURVE

Constants of the normal curve. The most important constants connected with the normal curve of error are the following (see Figure 3):

y_0 = maximum ordinate (where $x = 0$), or height at the "mode."

A = total area, from $x = -\infty$ to $x = +\infty$. If the curve is given by a finite number of equi-spaced ordinates, then, approximately, $A = N\cdot\Delta x$, where N = total length of the ordinates ("total population"), and Δx = distance between the ordinates ("class interval").

σ = "standard deviation" or "root-mean-square error" = abscissa of point of inflection, given by

$$\sigma = \sqrt{\frac{1}{A}\int_{-\infty}^{+\infty} x^2y\,dx}, \text{ or, approximately, } \sigma = \sqrt{\frac{1}{N}\Sigma(x^2y)}.$$

p = "probable error" = value of the abscissa such that the area from $x = -p$ to $x = +p$ is half the total area A. Here $p = (\rho\sqrt{2})\,\sigma = 0.674489749\,\sigma$, where $\rho = 0.476936276\cdots$ is a number defined by the equation

$$\frac{2}{\sqrt{\pi}}\int_0^{\rho} e^{-t^2}dt = \tfrac{1}{2}.$$

Less important constants are:

$1/h$ = "modulus" $= \sigma\sqrt{2}$. Here $h = 1/(\sigma\sqrt{2})$ is called the "measure of precision."

η = "mean absolute error" $= \sigma\sqrt{2/\pi} = 0.797884561\,\sigma$.

Here $\eta = \dfrac{2}{A}\displaystyle\int_0^\infty xy\,dx$, or, approximately, $\eta = \dfrac{1}{N}\Sigma(|x|\cdot y)$.

Note that σ, p, $1/h$, and η are quantities of the same kind as x, their order of magnitude being indicated in Figure 3.

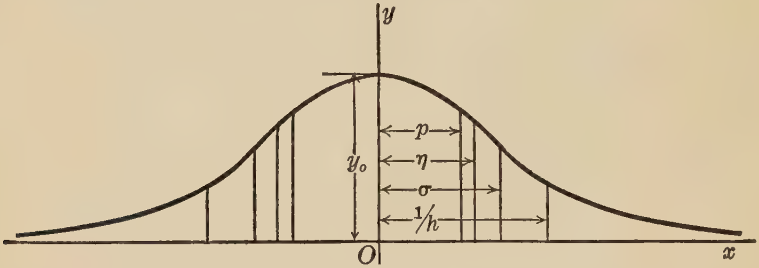


FIG. 3

Relations between the constants are shown in the following table (in which $\rho = 0.47694$, as defined above):

y_0	$= \dfrac{1}{\sqrt{2\pi}}\dfrac{A}{\sigma}$ $= 0.39894\dfrac{A}{\sigma}$	$= \dfrac{\rho}{\sqrt{\pi}}\dfrac{A}{p}$ $= 0.26908\dfrac{A}{p}$	$= \dfrac{1}{\sqrt{\pi}}Ah$ $= 0.56419\,Ah$	$= \dfrac{1}{\pi}\dfrac{A}{\eta}$ $= 0.31831\dfrac{A}{\eta}$
A	$= (\sqrt{2\pi})y_0\sigma$ $= 2.50663\,y_0\sigma$	$= \dfrac{\sqrt{\pi}}{\rho}y_0p$ $= 3.71633\,y_0p$	$= \sqrt{\pi}\dfrac{y_0}{h}$ $= 1.77245\dfrac{y_0}{h}$	$= \pi y_0\eta$ $= 3.14159\,y_0\eta$
σ	$= \sigma$	$= \dfrac{p}{\rho\sqrt{2}}$ $= 1.48260\,p$	$= \dfrac{1}{\sqrt{2}}\dfrac{1}{h}$ $= 0.70711\dfrac{1}{h}$	$= \sqrt{\dfrac{\pi}{2}}\eta$ $= 1.25331\,\eta$
p	$= (\rho\sqrt{2})\sigma$ $= 0.67449\,\sigma$	$= p$	$= \rho(1/h)$ $= 0.47694\dfrac{1}{h}$	$= (\rho\sqrt{\pi})\eta$ $= 0.845348\,\eta$
$\dfrac{1}{h}$	$= (\sqrt{2})\sigma$ $= 1.41421\,\sigma$	$= (1/\rho)p$ $= 2.09672\,p$	$= 1/h$	$= (\sqrt{\pi})\eta$ $= 1.77245\,\eta$
η	$= \left(\sqrt{\dfrac{2}{\pi}}\right)\sigma$ $= 0.79788\,\sigma$	$= \dfrac{p}{\rho\sqrt{\pi}}$ $= 1.18295\,p$	$= \dfrac{1}{\sqrt{\pi}}\dfrac{1}{h}$ $= 0.56419\dfrac{1}{h}$	$= \eta$

Equation of the curve. The equation of the normal curve may be written in seven ways, differing merely in respect to the choice of constants [indicated in square brackets] which it is desired to reduce to unity. (1) and (2) are in terms of the standard deviation, σ; (3) and (4) in terms of the probable error, p; (5) and (6) in terms of the modulus, $1/h$.

(1) $[\sigma, A]$	$\frac{y}{A/\sigma} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}$ $y_0 = 0.39894\,(A/\sigma)$	(2) $[\sigma, y_0]$	$\frac{y}{y_0} = e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}$ $A = 2.50663\,(\sigma y_0)$
(3) $[p, A]$	$\frac{y}{A/p} = \frac{\rho}{\sqrt{\pi}} e^{-\rho^2\left(\frac{x}{p}\right)^2}$ $y_0 = 0.26908\,(A/p)$	(4) $[p, y_0]$	$\frac{y}{y_0} = e^{-\rho^2\left(\frac{x}{p}\right)^2}$ $A = 3.71633\,(p y_0)$
(5) $[h, A]$	$\frac{y}{Ah} = \frac{1}{\sqrt{\pi}} e^{-(hx)^2}$ $y_0 = 0.56419\,(Ah)$	(6) $[h, y_0]$	$\frac{y}{y_0} = e^{-(hx)^2}$ $A = 1.77245\,(y_0/h)$
(7) $[A, y_0]$	$\frac{y}{y_0} = e^{-\pi\left(\frac{x}{A/y_0}\right)^2}$		

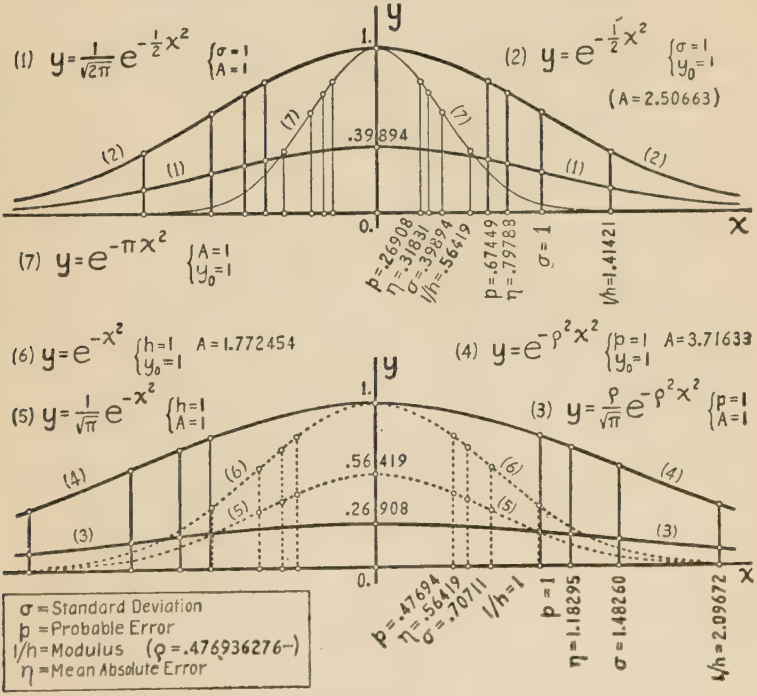


FIG. 4

The table on pages 209–16 is based on equation (1), giving $\frac{y}{(A/\sigma)}$ in terms of $\frac{x}{\sigma}$.

A table based on equation (2), giving $\frac{y}{y_0}$ in terms of $\frac{x}{\sigma}$, is found in Yule, page 303, 1922 Edition.

The probability integral, α, is a function of x giving the area under the normal curve, from $-x$ to $+x$. It may be expressed in seven different forms, corresponding to the seven forms of the equation of the curve. The letters in [ ] indicate the quantities which may conveniently be put equal to unity. As above, $\rho = 0.47694$.

(1) $[\sigma, A]$	$\frac{\alpha}{A} = \frac{2}{\sqrt{2\pi}} \int_0^{x/\sigma} e^{-\frac{1}{2}t^2} dt$ $= \frac{2}{\sqrt{\pi}} \int_0^{\frac{1}{\sqrt{2}}\frac{x}{\sigma}} e^{-t^2} dt$	(2) $[\sigma, y_0]$	$\frac{\alpha}{\sigma y_0} = 2 \int_0^{x/\sigma} e^{-\frac{1}{2}t^2} dt$ $= 2\sqrt{2} \int_0^{\frac{1}{\sqrt{2}}\frac{x}{\sigma}} e^{-t^2} dt$
(3) $[p, A]$	$\frac{\alpha}{A} = \frac{2\rho}{\sqrt{\pi}} \int_0^{x/p} e^{-\rho^2 t^2} dt$ $= \frac{2}{\sqrt{\pi}} \int_0^{\rho x/p} e^{-t^2} dt$	(4) $[p, y_0]$	$\frac{\alpha}{p y_0} = 2 \int_0^{x/p} e^{-\rho^2 t^2} dt$ $= \frac{2}{\rho} \int_0^{\rho x/p} e^{-t^2} dt$
(5) $[h, A]$	$\frac{\alpha}{A} = \frac{2}{\sqrt{\pi}} \int_0^{hx} e^{-t^2} dt$	(6) $[h, y_0]$	$\frac{\alpha}{y_0/h} = 2 \int_0^{hx} e^{-t^2} dt$
(7) $[A, y_0]$	$\frac{\alpha}{A} = 2 \int_0^{\frac{x}{A/y_0}} e^{-\pi t^2} dt = \frac{2}{\sqrt{\pi}} \int_0^{\sqrt{\pi}\frac{x}{A/y_0}} e^{-t^2} dt$		

The table on pages 209–16 is based on equation (1), giving $\frac{1}{2}\frac{\alpha}{A}$ in terms of $t = \frac{x}{\sigma}$. Sheppard's Table, in *Pearson's Tables*, gives $\frac{1}{2} + \frac{1}{2}\frac{\alpha}{A}$ in terms of $\frac{x}{\sigma}$. Encke's Table, reproduced in Wright and Hayford, Brunt, etc., gives $\frac{\alpha}{A}$ in terms of $\frac{x}{p}$, based on (3). Oppolzer's Table, based on (6), gives $\frac{1}{2}\frac{\alpha}{y_0/h}$ in terms of hx.

Burgess's Table, reproduced in most of the books, is based on (5), giving $\frac{\alpha}{A}$ in terms of hx.

Approximate expressions for the probability integral.

Let A = total area under the normal curve
and α = area from $-x$ to $+x$. Then

$$\frac{\alpha}{A} = \frac{2}{\sqrt{\pi}}\int_0^{hx} e^{-t^2}dt = \frac{2}{\sqrt{2\pi}}\int_0^{x/\sigma} e^{-\frac{1}{2}t^2}dt, \text{ where } A = (\sqrt{\pi})y_0\left(\frac{1}{h}\right) = (\sqrt{2\pi})y_0\sigma.$$

For small values of x:

$$\frac{\alpha}{A} = \frac{2}{\sqrt{\pi}}\left\{(hx) - \frac{(hx)^3}{1!3} + \frac{(hx)^5}{2!5} - \frac{(hx)^7}{3!7} + \frac{(hx)^9}{4!9} - \cdots\right\}.$$

$$\frac{\alpha}{A} = \frac{2}{\sqrt{2\pi}}\left\{(x/\sigma) - \frac{(x/\sigma)^3}{2\cdot 1!3} + \frac{(x/\sigma)^5}{2^2\cdot 2!5} - \frac{(x/\sigma)^7}{2^3\cdot 3!7} + \frac{(x/\sigma)^9}{2^4\cdot 4!9} - \cdots\right\}.$$

(Series convergent; error less than last term retained.)

For large values of x:

$$\frac{\alpha}{A} = 1 - 2\frac{e^{-(hx)^2}}{(hx)2\sqrt{\pi}}\left\{1 - \frac{1}{2(hx)^2} + \frac{1\cdot 3}{2^2(hx)^4} - \frac{1\cdot 3\cdot 5}{2^3(hx)^6} + \frac{1\cdot 3\cdot 5\cdot 7}{2^4(hx)^8} - \cdots\right\}.$$

$$\frac{\alpha}{A} = 1 - 2\frac{e^{-\frac{1}{2}(x/\sigma)^2}}{(x/\sigma)\sqrt{2\pi}}\left\{1 - \frac{1}{(x/\sigma)^2} + \frac{1\cdot 3}{(x/\sigma)^4} - \frac{1\cdot 3\cdot 5}{(x/\sigma)^6} + \frac{1\cdot 3\cdot 5\cdot 7}{(x/\sigma)^8} - \cdots\right\}.$$

(Series " asymptotic " or " semi-convergent "; but error is less than last term retained.)

Moments of the area of the normal curve, about the y-axis. The nth moment, μ_n, is defined by $\mu_n = \int_{-\infty}^{+\infty} x^n y dx$, where ydx is an element of area, and x the distance of that element from the y-axis.

If n is odd, $\mu_1 = \mu_3 = \mu_5 = \cdots = 0$. If n is even, we have $\mu_2 = A\sigma^2$, $\mu_4 = 1\cdot 3\,A\sigma^4$, $\mu_6 = 1\cdot 3\cdot 5\,A\sigma^6$, $\mu_8 = 1\cdot 3\cdot 5\cdot 7\,A\sigma^8, \cdots$, where A is the total area under the curve, and σ is the standard deviation.

Note that $\mu_4/(\mu_2{}^2) = 3/A$, $\mu_6/(\mu_2{}^3) = 15/A^2$, $\mu_8/(\mu_2{}^4) = 105/A^3, \cdots$.

FUNDAMENTAL FORMULAS IN PROBABILITY

The most important elementary formulas in probability, in the form in which they are usually stated, are here collected for reference. The true meaning and scope of these formulas is still a matter of much controversy.[1]

Probability *a priori*. If n events are regarded as " equally likely " to happen, as far as we can judge on the basis of a given body of in-

[1] For a recent critique, with full references, see J. M. Keynes, *A Treatise on Probability.*

formation, and if m of these events are "favorable" while the rest are "unfavorable," then the ratio $p = \frac{m}{n}$ is called the "probability of success," while $q = 1 - p$ is the "probability of failure."

For example, consider the throw of a perfect die; as far as we can judge, any one of the six forces is as likely to turn up as any other; hence the probability of throwing a 2-spot is $\frac{1}{6}$. Again, in throwing two dice, there are 36 equally probable results, of which two (namely, 5, 6 and 6, 5) will yield a total of 11; hence the probability of throwing 11 with two dice is $\frac{1}{18}$.

Addition formula. If $e_1, e_2, \cdots$ are "mutually exclusive" events, and if their separate probabilities are $p_1, p_2, \cdots$, then the probability that *some one* of the events will happen is the sum of the separate probabilities: $p = p_1 + p_2 + \cdots$. For example, the probability of throwing a 1-spot or a 2-spot with one die is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

Multiplication formula. If $e_1, e_2, \cdots$ are "independent" events, with separate probabilities $p_1, p_2, \cdots$, then the probability that *all* of the events will happen at once is the product of the separate probabilities: $p = p_1p_2 \cdots$. For example, the probability of throwing a double 2-spot with two dice is $(\frac{1}{6})(\frac{1}{6}) = \frac{1}{36}$.

Mathematical expectation. If $p =$ the probability of winning an amount w, then pw is called the mathematical expectation of the player.

Bernoulli's theorem. This theorem consists of two parts. (*a*) If an event is capable of being repeated many times, and the *a priori* probability of success at each trial is always p, then m, the most probable number of successes in n trials, will be np when np is an integer; or in any case $np - q \gtreqqless m \gtreqqless np + p$ (where $q = 1 - p$).

(*b*) Further, the probability that the actual number of successes shall differ from the most probable number, pn, by less than a given amount c, is approximately

$$P_c \doteq \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt + \frac{e^{-x^2}}{\sqrt{2\pi npq}},$$

where $q = 1 - p$ and $x = c/\sqrt{2npq}$.

Probability *a posteriori*. Suppose an event has occurred as a result of one or another of several causes, $C_1, C_2, \cdots$. Let p_i be the probability that C_i is present (before the occurrence of the event). Let P_i be the probability that C_i, if present, would produce the event. Then the probability that C_i was the actual cause is

$$(p_i P_i)/\Sigma(p_i P_i).$$

QUADRATURE FORMULAS, FOR NUMERICAL INTEGRATION

The area under a curve. The definite integral, $A = \int_a^b f(x)dx$, represents the area under the curve $y = f(x)$ from $x = a$ to $x = b$ (that is, the area bounded by the curve, the axis of x, and the ordinates corresponding to $x = a$ and $x = b$; any area below the axis being taken as negative). The problem of quadrature is to compute this area (approximately) when a finite number of equally spaced ordinates are given. The best-known formulas for this purpose are listed below.

We suppose the interval from $x = a$ to $x = b$ is divided into n equal parts, of length h, so that $h = (b - a)/n$. Also let

$$y_0 = f(a),\ y_1 = f(a + h),\ y_2 = f(a + 2\,h), \cdots$$
$$\cdots y_{n-2} = f(b - 2\,h),\ y_{n-1} = f(b - h),\ y_n = f(b).$$

(1) **Trapezoidal rule:**

$$A = h\{\tfrac{1}{2}\,y_0 + y_1 + y_2 + \cdots + y_{n-2} + y_{n-1} + \tfrac{1}{2}\,y_n\}.$$

The error involved in using this rule, that is, the correctional term which would have to be added to make the rule exact, is $-\left[\frac{nh^3}{12}\right]f''(X)$, where the accents denote differentiation, and X is some unknown value of x between a and b.

(2) **Simpson's rule** (if n is a multiple of 2):

$$A = \tfrac{1}{3}\,h\{y_0 + (4\,y_1 + 2\,y_2) + (4\,y_3 + 2\,y_4) + \cdots$$
$$\cdots + (2\,y_{n-2} + 4\,y_{n-1}) + y_n\}.$$

Here the correctional term is $-\left[\frac{nh^5}{180}\right]f''''(X)$, where X is some unknown value of x between a and b.

The special case when $n = 2$ is called the *prismoid formula:*

$$A = (h/3)\{y_0 + 4\,y_1 + y_2\},\text{ with error } = -\,(h^5/90)\,f''''(X).$$

If $f(x)$ is a polynomial of not higher than the third degree, its fourth derivative will be zero, and the prismoid formula will be exact. Further, if $f(x)$ is a polynomial of the fourth degree, its fourth derivative will be constant, and exact results can be obtained by using the prismoid formula with the correctional term.

(3) **The "three-eighths" rule** (if n is a multiple of 3):

$$A = \tfrac{3}{8}\,h\{y_0 + (3\,y_1 + 3\,y_2 + 2\,y_3) + (3\,y_4 + 3\,y_5 + 2\,y_6) + \cdots$$
$$\cdots + (2\,y_{n-3} + 3\,y_{n-2} + 3\,y_{n-1}) + y_n\}.$$

The special case when $n = 3$ gives

$$A = \tfrac{3}{8}\,h\{y_0 + 3\,y_1 + 3\,y_2 + y_3\}.$$

(4) **Weddle's rule** (if n is a multiple of 6):

$$\begin{aligned} A = \tfrac{3}{10}\,h\{y_0 &+ (5\,y_1 + y_2 + 6\,y_3 + y_4 + 5\,y_5 + 2\,y_6) \\ &+ (5\,y_7 + y_8 + 6\,y_9 + y_{10} + 5\,y_{11} + 2\,y_{12}) + \cdots \\ &+ (2\,y_{n-6} + 5\,y_{n-5} + y_{n-4} + 6\,y_{n-3} + y_{n-2} + 5\,y_{n-1}) + y_n\}. \end{aligned}$$

(5) **Sheppard's rule XI** (if n is a multiple of 12):

$$\begin{aligned} A &= A_1 + \tfrac{3}{5}(A_1 - A_2) - \tfrac{1}{5}(A_2 - A_3) + \tfrac{1}{35}(A_3 - A_4), \text{ where} \\ A_1 &= h(\tfrac{1}{2}\,y_0 + y_1 + y_2 + \cdots + y_{n-2} + y_{n-1} + \tfrac{1}{2}\,y_n), \\ A_2 &= 2\,h(\tfrac{1}{2}\,y_0 + y_2 + y_4 + \cdots + y_{n-4} + y_{n-2} + \tfrac{1}{2}\,y_n), \\ A_3 &= 3\,h(\tfrac{1}{2}\,y_0 + y_3 + y_6 + \cdots + y_{n-6} + y_{n-3} + \tfrac{1}{2}\,y_n), \\ A_4 &= 4\,h(\tfrac{1}{2}\,y_0 + y_4 + y_8 + \cdots + y_{n-8} + y_{n-4} + \tfrac{1}{2}\,y_n). \end{aligned}$$

(6) **The Euler-Maclaurin formula:**

$$A = h\{\tfrac{1}{2}y_0 + y_1 + y_2 + \cdots + y_{n-2} + y_{n-1} + \tfrac{1}{2}y_n\} - \frac{B_1h^2}{2\,!}[f'(b) - f'(a)]$$

$$+ \frac{B_3h^4}{4\,!}[f'''(b) - f'''(a)] + \cdots + \frac{(-1)^kB_{2k-1}h^{2k}}{(2\,k)\,!}[f^{(2k-1)}(b) - f^{(2k-1)}(a)] + R.$$

Here $R = \dfrac{(-1)^{k+1}B_{2k+1}h^{2k+2}}{(2\,k+2)\,!}\,nf^{(2k+2)}(X)$,

where X is some unknown number between a and b; and $B_1 = \frac{1}{6}$, $B_3 = \frac{1}{30}$, $B_5 = \frac{1}{42}$, $\cdots$ are the Bernoulli numbers.

INFINITE SERIES

Taylor's theorem. If a function $f(x)$ has derivatives of all orders at a point $x = a$, then for any value of x sufficiently near the value $x = a$, the function may be expanded into a power series arranged according to ascending powers of $x - a$, as follows:

$$f(x) = f(a) + \frac{f'(a)}{1!}(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \cdots + \frac{f^n(a)}{n!}(x - a)^n + R_{n+1},$$

where the remainder, R_{n+1}, lies between the largest and smallest values of $\dfrac{f^{n+1}(\xi)}{(n+1)\,!}(x - a)^{n+1}$ for values of ξ between a and x.

If the remainder, R_{n+1}, is small, the first few terms of the series give a good approximation to the value of the function.

Maclaurin's theorem is a special case of Taylor's theorem when $a = 0$. A few important expansions are as follows:

$$(1+x)^n = 1 + nx + \frac{n(n-1)}{2!}x^2 + \frac{n(n-1)(n-2)}{3!}x^3 + \cdots \qquad -1 < x < +1$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \qquad -\infty < x < +\infty$$

$$\log_e(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \qquad -1 < x < +1$$

$$\log_e\left(\frac{1+x}{1-x}\right) = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \frac{x^7}{7} + \cdots\right) \qquad -1 < x < +1$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \qquad -\infty < x < +\infty$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \cdots \qquad -\infty < x < +\infty$$

$$\tan x = x + \frac{x^3}{3} + \frac{2}{15}x^5 + \cdots \qquad -\frac{\pi}{2} < x < +\frac{\pi}{2}$$

$$\frac{x}{e^x - 1} = 1 - \frac{x}{2} + \frac{B_1 x^2}{2!} - \frac{B_3 x^4}{4!} + \frac{B_5 x^6}{6!} - \frac{B_7 x^8}{8!} + \cdots \qquad x < 2\pi$$

where $B_1 = \frac{1}{6}$, $B_3 = \frac{1}{30}$, $\cdots$ are the Bernoulli numbers.

Fourier's series. While a Taylor's series approximates a function near a particular point, $x = a$, a Fourier's series gives an approximation over a whole range.

If $f(x)$ is finite, and has only a finite number of discontinuities and only a finite number of maxima and minima, in an interval from $x = -c$ to $x = c$, then for any value of x between $-c$ and $+c$

$$f(x) = \frac{1}{2}a_0 + a_1 \cos\frac{\pi x}{c} + a_2 \cos\frac{2\pi x}{c} + a_3 \cos\frac{3\pi x}{c} + \cdots$$
$$+ b_1 \sin\frac{\pi x}{c} + b_2 \sin\frac{2\pi x}{c} + b_3 \cos\frac{3\pi x}{c} + \cdots,$$

where $a_n = \frac{1}{c}\int_{-c}^{c} f(t)\cos\frac{n\pi t}{c}\,dt$ and $b_n = \frac{1}{c}\int_{-c}^{c} f(t)\sin\frac{n\pi t}{c}\,dt.$

We note that $\int_0^\pi \sin^2 mx dx = \frac{\pi}{2}$ and $\int_0^\pi \cos^2 mx dx = \frac{\pi}{2}$, while if k is different from m,

$$\int_0^\pi \sin kx \cdot \sin mx \cdot dx = 0 \text{ and } \int_0^\pi \cos kx \cdot \cos mx \cdot dx = 0.$$

CHAPTER II

FREQUENCY DISTRIBUTIONS AVERAGES AND MEASURES OF DISPERSION · (ELEMENTARY METHODS)

By H. L. RIETZ

INTRODUCTION

The material of statistics. Items and collections of items may be regarded as the material about which the science of statistics is being developed. Each item may be simply the record of the presence of a certain quality in an individual; for example, it may indicate the color, sex, or occupation of an individual. Each item may be the result of counting or enumeration; for example, the number of children in a family or in a school, the population of a state, the number of births, deaths, or accidents. Each item may be a measurement or an estimate: for example, height, weight, or annual income of an individual; monthly pig iron production, interest rates, commodity prices, and so on.

Variables. Inequality among the items is a property of a statistical collection. On account of this property, a symbol used to represent the magnitudes of the items of a set may be appropriately called a **variable.**

We shall find it convenient to recognize two classes of variables: discrete and continuous.

A **discrete** variable is one whose values differ by assigned steps, often by unity; for example, the number of children in a family, the number of rows of kernels on an ear of corn.

A **continuous** variable is one whose values may differ by amounts which are indefinitely small; for example, the weight of a man, the temperature at a place.

FREQUENCY DISTRIBUTION[1]

A **frequency distribution** is an arrangement which shows the frequencies of the values of a variable in ordered classes. We may exhibit frequency distributions as follows:

[1] For early writings on frequency theory, see Ellis, *Tran. of Camb. Phil. Soc.*, vols. 8, 9 (1843–44); Venn, *Logic of Chance* (1866).

Example 1.

Number of rows of kernels on ears of corn	10	12	14	16	18	20	22	24
Frequencies	1	16	109	241	235	116	41	10

Example 2.

Weight in pounds of men aged 20–24, height 5′ 3″	100	105	110	115	120	125	130	135	140	145	150	155	160	165	170
Frequencies[1]	3	6	23	48	56	86	78	45	30	22	11	3	3	1	1

Example 3.

Number of "alpha-particles" radiated from a disk in one eighth of a minute	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Frequencies[2]	57	203	383	525	532	408	273	139	45	27	10	4	0	1	1

The first of these distributions is clearly with respect to a discrete variable with its values equal to even integers only. The second is with respect to a continuous variable. The third is with respect to a discrete variable with integral values only.

Class interval and class mark. A **class interval** is an interval which sets bounds to a class of a frequency distribution; for example, 102.5 to 107.5, 107.5 to 112.5 are class intervals for weight in the above example 2. A **class mark** is the value which is represented by the mid-value of the class interval; for example, 100, 105, 110, . . . are class marks in example 2.

With a discrete variable, the exact values of the variable are usually given by class marks. With a continuous variable, all the values that fall within a given class interval are for certain purposes grouped at the class mark as a convenient approximation.

The number of values of a variable within a class interval is called a **class frequency,** for example, 1, 16, 109, 241, 235, 116, 41, 10 are the class frequencies of example 1.

Frequency polygon. With a convenient horizontal scale for weight in pounds and a vertical scale for class frequencies, we plot in Figure 5 the class frequencies of example 2 as ordinates at the class marks 100, 105, 110, . . . shown below the base line. The **frequency polygon** (Fig. 5) is obtained by joining these points by straight lines. Each end point is joined to the base at the next class mark to close the polygon.

Frequency rectangles. Construct rectangles such as $ABCD$ (Fig. 5) with class intervals as bases, and with altitudes equal to the ordinates

[1] *Medico-Actuarial Investigation,* vol. 1 (1912), p. 40.

[2] Rutherford and Geiger, "The Probability Variations in the Distribution of Alpha-Particles," *Phil. Mag.*, series 6, vol. 20 (1910), pp. 698–701.

of points such as are plotted in Figure 5. We call these rectangles the frequency rectangles.

The upper boundaries of these frequency rectangles, together with the vertical segments joining the ends of these upper bases, as shown in Figure 5, give a graph called the **histogram** of the distribution.

Significance of area of frequency rectangles. Let us define unit area as that of a rectangle whose base is a class interval and whose altitude represents a unit of frequency. Then the area of $ABCD$ (Fig. 5) is 48,

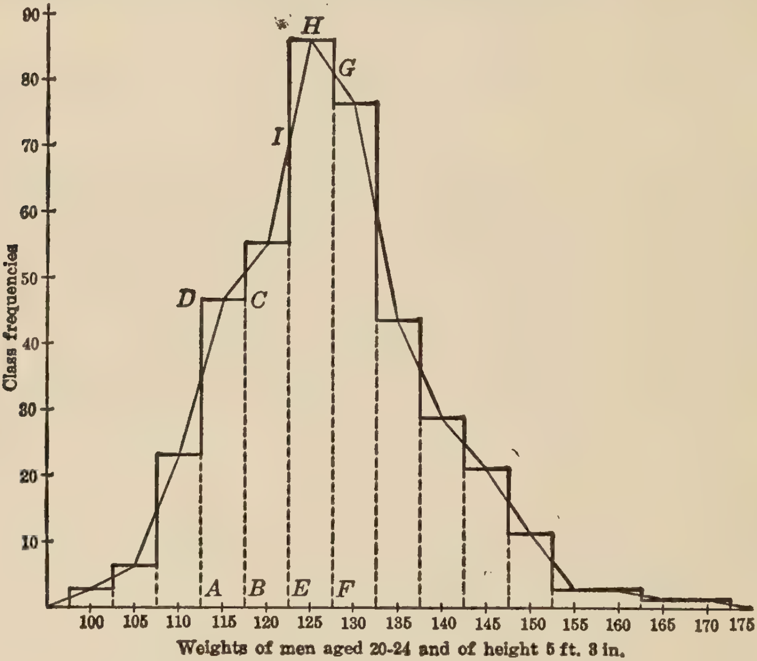


FIG. 5

and the total area of all the frequency rectangles is equal to the total frequency.

Frequency curve. A **frequency curve** may be regarded as an estimate of the limit that would probably be approached by a frequency polygon or histogram (Fig. 5) if the class intervals were made smaller and smaller, and the frequency N were at the same time increased without limit. The number of observations that will in the long run fall between assigned values a and b is proportional to the shaded area. Thus, if $y = f(x)$ is the frequency curve,

$$\int_a^b y dx$$

is the most probable frequency of observations in the interval a to b.

Sometimes the whole area under a frequency curve is taken as the unit of area. Then

$$\int_{-\infty}^{+\infty} ydx = 1$$

and an area represents a relative frequency or statistical probability.

Thus,

$$\int_{a}^{b} ydx$$

gives the probability that an observation taken at random from a set will fall into the interval a to b.

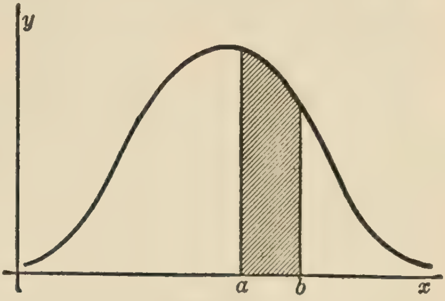


FIG. 6

A cumulative frequency distribution. Start at one end of a frequency distribution such as that in example 2 by recording the end class frequency, then form the sum of two class frequencies nearest the end, then of the three nearest the end, and so on. When the end value and these sums are arranged in the order in which they are obtained, the distribution is called a **cumulative frequency distribution.** For example, by thus adding frequencies in example 2, page 21, we obtain

Example 4.

Weight in pounds below	102.5	107.5	112.5	117.5	122.5	127.5	132.5	137.5	142.5	147.5	152.5	157.5	162.5	167.5	172.5
Cumulative frequencies .	3	9	32	80	136	222	300	345	375	197	408	411	414	415	416

as a cumulative frequency distribution. It shows the number of individuals whose weight does not exceed an assigned value.

Ogive. The graph of a cumulative frequency distribution on coördinate paper is called an **ogive.**

APPLICATION OF AVERAGES TO FREQUENCY DISTRIBUTIONS

In the description of a frequency distribution, we usually make use of certain averages already defined in Chapter I. In this connection, the averages may take on further meanings in the characterization of the distribution.

Graphical meaning of the arithmetic mean. In the graphical representation of a frequency distribution (Figs. 5 and 6), the arithmetic mean (AM) of all the values of a variable is the abscissa of the centroid of the total area under the frequency curve. From a sample set consisting of N values of the variable, we actually compute the abscissa of the centroid of the area of the frequency rectangles.

Computation of the arithmetic mean $\bar{X}$ when a frequency distribution is given.

Let $f_1, \ f_2, \cdots f_n$ be class frequencies,

$X_1, X_2, \cdots X_n$ corresponding class marks,

and $N = f_1 + f_2 \cdots + f_n$.

By definition, the arithmetic mean

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + \cdots + f_nX_n}{f_1 + f_2 + \cdots + f_n} = \frac{\Sigma fX}{\Sigma f} \tag{1}$$

$$= X_0 + \frac{f_1(X_1 - X_0) + f_2(X_2 - X_0) + \cdots + f_n(X_n - X_0)}{N}$$

$$= X_0 + \frac{1}{N}\Sigma f(X - X_0), \tag{2}$$

where X_0 is an arbitrary number.

The formula (2), with X_0 chosen as the class mark near the mean, is usually more convenient than (1) for numerical computation, when N is large.

The arithmetic mean is then

$$\bar{X} = X_0 + b, \tag{3}$$

where $b = \dfrac{f_1(X_1 - X_0) + f_2(X_2 - X_0) + \cdots + f_n(X_n - X_0)}{N}$

$$= \frac{1}{N}\Sigma f(X - X_0) \tag{4}$$

is the mean value of the deviations $X - X_0$ and may be regarded as the correction to be applied to the guess X_0 to get the mean. A form for the computation of $\bar{X}$ by this method is shown in example 5.

Weighted arithmetic mean. The weighted arithmetic mean of X_1, $X_2, \cdots X_n$ with weights $W_1, W_2, \cdots W_n$ is defined as

$$\bar{X} = \frac{W_1X_1 + W_2X_2 + \cdots + W_nX_n.}{W_1 + W_2 + \cdots + W_n} \tag{5}$$

It is obvious from (1) and (5) that $\bar{X}$ in (1) may be regarded as the arithmetic mean of class marks weighted with corresponding class frequencies.

Form for the calculation of the arithmetic mean for a given frequency distribution. Let X = class mark, f = frequency, X_0 = a selected class mark near the mean. Deviations $X - X_0$ are in units of a class interval.

Example 5. Finding the AM from the frequency distribution of example 2.

(1) X	(2) f	(3) $X - X_0$	(4) $f(X - X_0)$	
170	1	9	9	
165	1	8	8	
160	3	7	21	
155	3	6	18	
150	11	5	55	
145	22	4	88	
140	30	3	90	
135	45	2	90	
130	78	1	78	
125	86	0	0	$\overline{457}$
120	56	− 1	− 56	
115	48	− 2	− 96	
110	23	− 3	− 69	
105	6	− 4	− 24	
100	3	− 5	− 15	
	416			− 260
				+ 197

$$b = +\frac{197}{416} = 0.474 \text{ class interval} = 2.37 \text{ pounds.}$$

$$\bar{X} = 125 + 2.37 = 127.37 \text{ pounds.}$$

A good check on the accuracy of the computation consists in recalculating the mean after changing the origin of deviations by one class interval.

Geometric mean.[1] In terms of the symbols defined above, the geometric mean (GM) is

$$GM = \sqrt[N]{X_1^{f_1} X_2^{f_2} \cdots X_n^{f_n}}$$

or

$$\log GM = \frac{f_1 \log X_1 + f_2 \log X_2 + \cdots + f_n \log X_n}{N}. \tag{6}$$

In the right-hand member of (6), we have simply the arithmetic mean of the logarithms, log X_1, log X_2, $\cdots$, log X_n, in place of the arithmetic mean of the numbers given in (1). Thus, the geometric mean is simply the antilogarithm of the arithmetic mean of the logarithms of the numbers. The calculation of the geometric mean is thus reduced to the calculation of the arithmetic mean of the logarithms.

Harmonic mean.[2] Since the harmonic mean (p. 5) of a set of num-

[1] For uses of geometrical mean, see G. Udny Yule, *Introduction to Theory of Statistics* (1911), pp. 125–28.

[2] For the meaning of a certain useful weighted arithmetic mean and of a corresponding weighted harmonic mean to be used in measuring changes in the general price level, see Allyn A. Young, "The Measurement of Changes in the General Price Level," *The Quarterly Journal of Economics*, vol. 35 (1921), pp. 557–73.

bers is the reciprocal of the arithmetic mean of the reciprocals of the numbers, the calculation of the harmonic mean is at once reduced to that of the calculating of an arithmetic mean.

Graphical meaning of the median. In the graphical representation of a frequency distribution (Figs. 5 and 6), the median of a variable is the abscissa of points the vertical through which divides the total area under the frequency curve into two equal parts. What we actually compute for the median of a sample set of N values of the variable is the abscissa of a point the vertical through which divides the total area of the frequency rectangles into two equal parts.

Computation of the median when the frequency distribution is given. The calculation of the median when the frequency distribution is given usually involves interpolation by proportional parts. We shall find as an illustration the median of weight of men for which the distribution is given in example 2. The median weight of the 416 men is a weight such that 208 are below this value and 208 are above it.

From the cumulative frequencies of example 4, page 23, we note that the median falls into the class interval 122.5 to 127.5. Below 122.5 there are 136 cases. Below 127.5 there are 222 cases.

Hence, of the class frequency, 86, at 125, there are 72 below the median and 14 above it. Hence, by interpolation by proportional parts, the median M is

$$M = 122.5 + \frac{(72)(5)}{86} = 126.69 \text{ pounds.} \qquad (7)$$

The accuracy of the result depends upon the usual assumption in such interpolations that the distribution is uniform in the interval from 122.5 to 127.5.

The mode. In the graphical representation of a frequency distribution (Figs. 5 and 6), the mode is the abscissa of a maximum value of the frequency function. Experience has shown that a relatively large number of frequency distributions have only one mode, but some have two or more modes.

The accurate determination of the mode is not usually a simple matter and belongs to a later chapter in which we determine the equations of frequency curves (Chap. VII).

In the frequency distribution of weights (p. 21), it may be noted that the class frequencies increase up to 86 at 125 and then decrease. The mode is then probably in the neighborhood of 125, but all values from 122.5 to 127.5 were placed in the 125-pound class.

The point at which the frequency is most dense is the abscissa of the maximum on the frequency curve, and can be determined accurately only from the equation of the curve.

For a given grouping, the class mark of the maximal class frequencies is called the **empirical mode** to distinguish it from the mode defined above. Thus, 125 is the empirical mode in example 2.

An approximation [1] to the mode, measured from the lower limit of the modal class, is given by

$$x = -\frac{\Delta f_{-1}}{\Delta^2 f_{-1}},$$

where f is the frequency of the modal class, f_{-1} the frequency of the class next below, f_1 that of the class next above the modal class; and Δ and Δ^2 denote first and second differences.

For the frequency distribution of example 2,

$$\begin{array}{rcc}
f_{-1} = 56 & & \\
 & 30 & \\
f = 86 & & -38 \\
 & -8 & \\
f_1 = 78 & &
\end{array}$$

Then $x = \frac{30}{38} = .79$ class interval units measured from the lower limit of the class 125. Thus, the mode is

$$122.5 + 5(.79) = 126.45 \text{ pounds.}$$

As indicated by Pearson [2] this method should be used with caution because of the large probable error in the results.

DISPERSION OR VARIABILITY

Measures of dispersion. After finding an average of a set of observations, it is usually important to determine the extent to which the values are scattered from this average.

If the dispersion is to be measured by a single number, it is surely appropriate to use some kind of an average of deviations. The "root-mean-square of deviations" and the "mean of the absolute values" of deviations are the averages which have been much used for this purpose.

It is our purpose to show how these measures of dispersion may be most conveniently calculated, and to give a notion of their geometrical meanings in the description of a frequency distribution.

Standard deviation. The standard deviation is the square root of the arithmetic mean of the squares of deviations of values of the variable

[1] E. Czuber, *Die Statistischen Forschungsmethoden* (1921), pp. 71–72.

[2] *Biometrika*, vol. 1 (1902), p. 260.

from their arithmetic mean. The standard deviation is very commonly denoted by σ. Thus, with the notation of page 24,

$$\sigma^2 = \frac{f_1(X_1 - \overline{X})^2 + f_2(X_2 - \overline{X})^2 + \cdots + f_n(X_n - \overline{X})^2}{N} \tag{8}$$

$$= \frac{1}{N}\sum_{t=1}^{t=n} f_t(X_t - \overline{X})^2. \tag{9}$$

The calculation of σ from this formula is likely to be much more laborious than the calculation from the equivalent formula

$$\sigma^2 = \frac{f_1(X_1 - X_0)^2 + f_2(X_2 - X_0)^2 + \cdots + f_n(X_n - X_0)^2}{N} - b^2, \tag{10}$$

where X_0 is a selected class mark near the mean as defined on page 24 and $b = \overline{X} - X_0$.

Graphical meaning of σ in a normal distribution. In case we have what is known as a normal distribution (see p. 11), the standard deviation is one half the distance between the points of inflection on the frequency curve.

Form for the calculation of the standard deviation of a given frequency distribution. We use here the notation shown in the form on page 25. We add to the four columns of that form a fifth column headed $f(X - X_0)^2$. This column is obtained in an obvious manner from columns (3) and (4). We add also a check column (6) headed $f(X + 1 - X_0)^2$ formed in an obvious manner, by lowering the origin of deviations in column (3) by one class interval.

Example 6. Finding the standard deviation of the frequency distribution of example 2. The columns (1), (2), (3), (4) are simply copied from example 5.

(1)	(2)	(3)	(4)	(5)	(6)
X	f	$X - X_0$	$f(X - X_0)$	$f(X - X_0)^2$	$f(X + 1 - X_0)^2$
170	1	9	9	81	100
165	1	8	8	64	81
160	3	7	21	147	192
155	3	6	18	108	147
150	11	5	55	275	396
145	22	4	88	352	550
140	30	3	90	270	480
135	45	2	90	180	405
130	78	1	78	78	312
125	86	0			86
120	56	− 1	− 56	56	
115	48	− 2	− 96	192	48
110	23	− 3	− 69	207	92
105	6	− 4	− 24	96	54
100	3	− 5	− 15	76	48
	416		197	2181	2991

$$\begin{aligned}
\Sigma f(X - X_0)^2 &= 2181. \\
\frac{1}{N}\Sigma f(X - X_0)^2 &= 5.243 \\
b^2 &= \underline{.225} \\
\sigma^2 &= 5.018^{1} \\
\sigma &= 2.240 \text{ class intervals} \\
&= 11.20 \text{ pounds.}
\end{aligned}$$

Charlier Check

$$\begin{aligned}
\Sigma f(X + 1 - X_0)^2 &= \Sigma f(X - X_0)^2 + 2\,\Sigma f(X - X_0) + \Sigma f. \\
2991 &= 2181 + 2(197) + 416 \\
&= 2991.
\end{aligned}$$

Coefficient of variability. The ratio $C = \frac{\sigma}{\overline{X}}$ of the standard deviation to the arithmetic mean is called the coefficient of variability and is usually expressed as a percentage.

Mean or average deviation. The mean or average deviation of the values of a variable from an average is the arithmetic mean of the deviations treating them all as positive. The deviations may be taken from the arithmetic mean or from the median, but the mean deviation is least when the median is taken as the origin or zero point from which deviations are measured.

Graphical meaning of mean deviation of a normal distribution. When we have a normal distribution (see page 11), and the origin is at the arithmetic mean or median, the mean deviation is simply the abscissa of the centroid of the area under the right-hand half of the frequency curve. Furthermore, there is, in the case of a normal distribution, a simple relation between the mean deviation and standard deviation. Thus,

$$\text{Mean deviation} = .7979\,\sigma = .8\,\sigma \text{ roughly.} \tag{11}$$

Calculation of mean deviation from a given frequency distribution. We find first the sum of the absolute values of the deviations from the center of the class interval within which the average falls from which the deviations are to be taken, but we do not include in this the values in that class interval in which the average falls.

We should next make corrections to this sum for the distance of the average from the center of the class interval in which the average lies, and for the fact that the observations in this class interval usually lie

[1] A correction, known as Sheppard's correction, may be applied to this approximate value of σ^2 to correct for grouping into 5-pound classes. The method of making this refinement in the calculation will be given in Chapter VII. (See page 93.) The use of Sheppard's correction is restricted to cases in which the frequency curve has high order contact with the x-axis at the ends of the range.

partly above and partly below the average. We shall refer to these two corrections as correction (1) and correction (2)[1] respectively. To be more specific, let us assume that the deviations are measured from the median.

Correction (1). If N_a is the number of observations above and N_b the number below the class interval in which the median lies, and c is the distance in units of class intervals from the center of the class interval to the median, we have as correction (1),

$$(N_b - N_a)c.$$

Correction (2). Let N_m be the number of observations in the class interval in which the median lies. If we assume these N_m values uniformly distributed over the unit interval, we have, by proportional parts, $(.5 + c)N_m$ values below the median and $(.5 - c)N_m$ values above the median. With a uniform distribution over the unit interval, the sum of these deviations below the median would be

$$\frac{(.5 + c)^2 N_m}{2} \text{ and above the median } \frac{(.5 - c)^2 N_m}{2}.$$

That is, the part of the area N_m below the median ordinate is $(.5 + c)N_m$ and the AM of the deviations of its points from the median is $\frac{.5 + c}{2}$. The product of $(.5 + c)N_m$ and $\frac{.5 + c}{2}$ is the sum of the deviations below the median ordinate. Similarly, the part of the area N_m above the median ordinate is $(.5 - c)N_m$ and the AM of the deviations of its points from the median is $\frac{.5 - c}{2}$. The sum of the deviations of the N_m values is then

$$\frac{(.5 + c)^2}{2} N_m + \frac{(.5 - c)^2}{2} N_m = (.25 + c^2)N_m.$$

Let X and f have the same meanings as the computation on page 24, and let $X - X_0$ be the absolute value of $X - X_0$. For the frequency distribution of example 2, the median found in (7) is 126.68 and thus lies in the interval 122.5 to 127.5. In units of class intervals the median is

$$\frac{126.69 - 122.5}{5} = 0.838$$

of a unit above the lower bound of the interval and 0.338 of a unit above the center of the interval.

[1] A less accurate method of finding the mean deviation than that presented here is given in certain current textbooks. Cf. G. Udny Yule, *Introduction to the Theory of Statistics*, pp. 145–46; H. Secrist, *An Introduction to Statistical Methods*, p. 396.

Example 7. **Form for the calculation of the mean deviation from the median $M = 126.68$ pounds.**

X	f	$\lvert X - X_0 \rvert$	$f\lvert X - X_0 \rvert$
170	1	9	9
165	1	8	8
160	3	7	21
155	3	6	18
150	11	5	55
145	22	4	88
140	30	3	90
135	45	2	90
130	78	1	78
125	86	0	
120	56	1	56
115	48	2	96
110	23	3	69
105	6	4	24
100	3	5	15
	416		717

Correction (1): $(N_b - N_a)c = (136 - 194)(.338) = -\ 19.6$

Correction (2): $(.25 + c^2)N_m = (.25 + .114)86 = +\ 31.3$

Sum of all deviations $= 728.7$

$$\text{Mean deviation} = \frac{728.7}{416} = 1.750 \text{ class intervals}$$
$$= 8.75 \text{ pounds.}$$

Quartile deviation — Semi-interquartile range. The lower quartile Q_1 is such a value of the variable in a frequency distribution that one quarter of the total frequency is below Q_1, and three quarters above it. The upper quartile Q_3 is such a value that three quarters of the total frequency is below Q_3 and one quarter above it.

The quartile deviation or semi-interquartile range is

$$\frac{Q_3 - Q_1}{2}. \tag{12}$$

The quartiles, like the median, may be calculated from the frequency distribution by interpolation. Thus, in example 7, the value with $\frac{416}{4} = 104$ values below it is in the interval 117.5 to 122.5. Below the lower bound of this interval, there is a frequency 80. Hence, by proportional parts, Q_1 is above this boundary by $\dfrac{104 - 80}{56} \times 5 = 2.14$.

Hence, we have $\qquad Q_1 = 117.5 + 2.14 = 119.64.$

Similarly, we have $\qquad Q_3 = 132.5 + \dfrac{312 - 300}{45} \times 5 = 133.83.$

Formulas for probable errors in certain averages. The meaning of the probable error in a statistical result obtained from N values of a variable is presented in a later chapter,[1] but for convenience we give here formulas for probable errors of certain averages and functions of averages.

Let PE be the probable error, σ the standard deviation, C the coefficient of variability, the subscript of PE the average or other statistical result whose probable error is to be given. For a normal[2] distribution,

$$(PE)_{AM} = \pm .6745 \frac{\sigma}{\sqrt{N}},$$

$$(PE)_{\sigma} = \pm .6745 \frac{\sigma}{\sqrt{2\,N}} = \pm .4769 \frac{\sigma}{\sqrt{N}}$$

$$(PE)_{median} = \pm .8454 \frac{\sigma}{\sqrt{N}},$$

$$(PE)_{quartile} = \pm .9191 \frac{\sigma}{\sqrt{N}},$$

$$(PE)_{semi\text{-}interquartile\ range} = \pm .5306 \frac{\sigma}{\sqrt{N}}.$$

For a normal distribution,

$$(PE)_C = \pm .6745 \frac{C}{\sqrt{2\,N}}\left[1 + 2\left(\frac{C}{100}\right)^2\right]^{\frac{1}{2}}$$

$$= \pm .4769 \frac{C}{\sqrt{N}}\left[1 + 2\left(\frac{C}{100}\right)^2\right]^{\frac{1}{2}}.$$

The appropriate average to use. The selection of the average which is most appropriate for describing a distribution of values depends much on the nature of the distribution and on the purpose for which the average is obtained. For example, if we wish the average cost of a set of N articles purchased at various prices, the most obvious and usual purpose would be to find that equal price per article which would give the same total cost as the given variable prices. Clearly, the arithmetic mean satisfies this condition.

Next, if we wish the average annual increase in population for ten years when we have given the rates of gain of each of the ten years, the main purpose would be to get the equal rates of gain r which would lead to the same result in ten years as the variable rates $r_1, r_1, \cdots r_{10}$. Clearly, we must then find the geometric mean $1 + r$ in the equation

$$(1 + r)^{10} = (1 + r_1)(1 + r_2) \cdots (1 + r_{10}).$$

[1] Chapter V, "Random Sampling."

[2] For meaning of a normal distribution, see Chapter VII.

Again, if we require for our purpose the middle value of a group, the median is obviously the average we should find. In another situation, we may require the most frequent value. Then the mode is obviously the average we should find.

While it is thus fairly obvious in some cases which average to use, it must be granted that for other cases, it is not at all clear, from the nature of the data or distribution, which average will serve the purpose best. In such cases, more than one average may well be used. For certain types of distribution, two or more averages give a much better description than any single average.

If various averages have been computed, but only a single one is to be reported, it appears reasonable that some preference should be given to the average that has the smallest probable error.

Skewness. The character of a frequency curve (Fig. 6) with respect to symmetry or skewness about the maximal ordinate is very commonly involved in an elementary description of a frequency distribution.

It is important, therefore, that we should have some means of measuring skewness. In a symmetrical distribution, the arithmetic mean, mode, and median coincide. In a skew distribution they do not coincide.

Since skewness relates to the shape rather than to the size of the curve, it seems appropriate to use a ratio as a measure of skewness. The ratio

$$\text{skewness} = \frac{\text{arithmetic mean} - \text{mode}}{\text{standard deviation}}$$

suggested by Karl Pearson has been much used to measure skewness. Skewness is also indicated when the quartiles, or pairs of deciles, are not equidistant from the median. With this fact as a basis, a measure

$$\text{skewness} = \frac{\text{upper quartile} + \text{lower quartile} - 2\ \text{median}}{\text{upper quartile} - \text{lower quartile}}$$

is sometimes used to measure skewness.

The elementary methods presented in this chapter for describing frequency distributions by means of averages and measures of dispersion characterize only certain features of the distribution. These methods, as well as the simplest freehand graphical representations of frequency curves, point to the desirability of more complete descriptions of the distributions based on interpolation theory and curve fitting. Chapters III, IV, and VII deal with these topics.

CHAPTER III

INTERPOLATION, SUMMATION, AND GRADUATION

BY JAMES W. GLOVER

UNDERLYING IDEAS OF INTERPOLATION

Interpolation by differences. It is often required to interpolate additional ordinates in a given series of statistical or functional values represented by a set of ordinates. These interpolated values must of course closely approximate the true values of the ordinates when they are analytical functions of a variable abscissa. The interpolated values are usually assumed to lie on a parabolic curve passing through the ends of the given ordinates. For example, u_0, u_1, u_2, u_3, determine uniquely the parabola

$$u_x = a_0 + a_1x + a_2x^2 + a_3x^3 \qquad (1)$$

of the third degree. The coefficients a_0, a_1, a_2, and a_3, are functions of the given ordinates and their differences. When x takes the values 0, 1, 2, 3, the function u_x takes the values u_0, u_1, u_2, u_3, respectively. If x is given a value between 0 and 1, the function u_x takes the value assigned to the corresponding interpolated ordinate.

The coefficients in (1) could be determined by substituting and solving the equations

$$\begin{aligned} u_0 &= a_0, \\ u_1 &= a_0 + a_1 + a_2 + a_3, \\ u_2 &= a_0 + 2\,a_1 + 4\,a_2 + 8\,a_3, \\ u_3 &= a_0 + 3\,a_1 + 9\,a_2 + 27\,a_3. \end{aligned} \qquad (2)$$

However, by employing the methods of finite differences expressions for u_x are obtained with less labor.

Let u_x denote a function of the variable x and consider a succession of equidistant ordinates, that is, separated by equal intervals (taken as unit intervals) along the x-axis:

$$\cdots u_{-3},\ u_{-2},\ u_{-1},\ u_0,\ u_1,\ u_2,\ u_3, \cdots \qquad (3)$$

Then by definition the successive differences are

$$\begin{array}{lll} \Delta u_{-3} = u_{-2} - u_{-3}, & \Delta u_{-2} = u_{-1} - u_{-2}, \cdots & \Delta u_2 = u_3 - u_2, \cdots \\ \Delta^2 u_{-3} = \Delta u_{-2} - \Delta u_{-3}, & \Delta^2 u_{-2} = \Delta u_{-1} - \Delta u_{-2}, \cdots & \Delta^2 u_2 = \Delta u_3 - \Delta u_2, \cdots \end{array} \qquad (4)$$

and so on.

Horizontal difference table. Having given a succession of values of a function, the process of differencing may be repeated and a horizontal table of differences is developed as follows:

TABLE I. HORIZONTAL DIFFERENCE TABLE

x	u_x	Δu_x	$\Delta^2 u_x$	$\Delta^3 u_x$	$\Delta^4 u_x$	$\Delta^5 u_x$	$\Delta^6 u_x$
-3	u_{-3}	Δu_{-3}	$\Delta^2 u_{-3}$	$\Delta^3 u_{-3}$	$\Delta^4 u_{-3}$	$\Delta^5 u_{-3}$	$\Delta^6 u_{-3}$
-2	u_{-2}	Δu_{-2}	$\Delta^2 u_{-2}$	$\Delta^3 u_{-2}$	$\Delta^4 u_{-2}$	$\Delta^5 u_{-2}$	
-1	u_{-1}	Δu_{-1}	$\Delta^2 u_{-1}$	$\Delta^3 u_{-1}$	$\Delta^4 u_{-1}$		
0	u_0	Δu_0	$\Delta^2 u_0$	$\Delta^3 u_0$			
1	u_1	Δu_1	$\Delta^2 u_1$				
2	u_2	Δu_2					
3	u_3						

The column headed x gives the values of the variable x, sometimes called the **argument.** The remaining columns give the values of the function and successive differences. The differences in the same row with the function have the same subscript as the function and are called the **leading differences** of that value of the function.

Fundamental relations in difference table. The truth of the following statements is evident from the construction of the difference table.

(*a*) The columns to the right of any given column contain the successive differences of that column.

(*b*) The first column to the left of any column, when differenced, will produce that column; the second column to the left when differenced twice, will produce that column, and so on.

If we define $\Delta^{-1}u_x$ as a function whose first difference is u_x, then the columns to the left of any column are (apart from arbitrary constants and periodic functions of the variable) the $\Delta^{-1}u_x$, $\Delta^{-2}u_x$, $\Delta^{-3}u_x$, and so on, of that column.

(*c*) If a column is inverted and the successive differences formed, the even differences are not affected and the odd differences are changed in sign only.

(*d*) The sum of any two successive differences in any row in a horizontal difference table is equal to the difference in the next row directly under the first difference. In symbols:

$$\Delta^k u_x + \Delta^{k+1} u_x = \Delta^k u_{x+1}. \tag{5}$$

These three differences lie in pairs in a row, in a column, and in a diagonal.

(*e*) The sum of a series of successive terms in any column of a horizontal difference table is equal to the following term in the next row and in the column to the left minus the term in the same row with the first term and in the column to the left. In symbols:

$$\sum_{x=a}^{x=s} u_x = \Delta^{-1}u_{s+1} - \Delta^{-1}u_a = \Delta^{-1}u_x \Big|_{x=a}^{x=s+1} \tag{6}$$

INTERPOLATION FORMULAS

Newton's interpolation formula. It is possible to express the function u_x in terms of the function and leading differences of any row. For example,

$$u_x = u_0 + x\Delta u_0 + x_2\Delta^2 u_0 + x_3\Delta^3 u_0 + x_4\Delta^4 u_0 + x_5\Delta^5 u_0 + \cdots, \tag{7}$$

where

$$x_k = \frac{x(x-1)\cdots(x-k+1)}{\underline{|k}}. \tag{8}$$

In terms of the horizontal difference table, the values in any row are employed to obtain the value in the first column corresponding to the argument x. This is known as Newton's interpolation formula.

Example 1. The values of

$$u_t = \frac{1}{\sqrt{2\pi}}\int_0^t e^{-t^2/2}\,dt, \tag{9}$$

the area under the normal curve of error to the right of the origin, are given at intervals of $\frac{1}{2}$ from $t = 0$ to $t = 4$ as follows:

t	u_t	t	u_t	t	u_t
.00	.00000	1.50	.43319	3.00	.49865
.50	.19146	2.00	.47725	3.50	.49977
1.00	.34134	2.50	.49379	4.00	.49997

Find the value of the area under the curve by Newton's interpolation formula when $t = 1.22$.

The horizontal difference table, with the intervals between equidistant ordinates taken as unit intervals, assumes the following form:

Table II

t	x	u_x	Δu_x	$\Delta^2 u_x$	$\Delta^3 u_x$	$\Delta^4 u_x$
.00	−2	.00000	.19146	−.04158	−.01645	.02670
.50	−1	.19146	.14988	−.05803	.01024	.01003
1.00	0	.34134	.09185	−.04779	.02027	−.00444
1.50	1	.43319	.04406	−.02752	.01584	
2.00	2	.47725	.01654	−.01168		
2.50	3	.49379	.00486			
3.00	4	.49865				

Substitution in (7) gives

$$\begin{aligned} u_{.44} &= .34134 + .44(.09185) + (-.1232)(-.04779) \\ &\qquad + .06406(.02027) + (-.041)(-.00444) \\ &= .34134 + .04041 + .00589 + .00130 + .00018 = .38912. \end{aligned}$$

The true value is .38877, hence the error is .00035.

Diagonal difference table. By (5) any term in the horizontal difference table can be expressed in terms of two differences in the row above, or the column to the left, or the diagonal below. In this manner, Newton's formula can be transformed by successive steps into other forms. These steps will perhaps be better understood by writing the **horizontal** difference table in the **diagonal** difference table form.

TABLE III. DIAGONAL DIFFERENCE TABLE

x	u_x	Δu_x	$\Delta^2 u_x$	$\Delta^3 u_x$	$\Delta^4 u_x$	$\Delta^5 u_x$	$\Delta^6 u_x$
-3	u_{-3}						
		Δu_{-3}					
-2	u_{-2}		$\Delta^2 u_{-3}$				
		Δu_{-2}		$\Delta^3 u_{-3}$			
-1	u_{-1}		$\Delta^2 u_{-2}$		$\Delta^4 u_{-3}$		
		Δu_{-1}		$\Delta^3 u_{-2}$		$\Delta^5 u_{-3}$	
0	u_0		$\Delta^2 u_{-1}$		$\Delta^4 u_{-2}$		$\Delta^6 u_{-3}$
		Δu_0		$\Delta^3 u_{-1}$		$\Delta^5 u_{-2}$	
1	u_1		$\Delta^2 u_0$		$\Delta^4 u_{-1}$		
		Δu_1		$\Delta^3 u_0$			
2	u_2		$\Delta^2 u_1$				
		Δu_2					
3	u_3						

The diagonal table may be thought of as obtained from the horizontal table by turning the line of leading differences in any row about the function at the beginning of that row as a pivot through a downward angle until the even differences lie in the first, second, and so on, rows, respectively, below their original positions. In the new positions the leading differences of a function lie in the downward diagonal setting out from the function. Accordingly Newton's interpolation formula for u_x expresses it in terms lying on a downward diagonal in the diagonal difference table.

The interpolation formulas of Gauss and Stirling's central difference interpolation formula. By employing (5),

$\Delta^2 u_0$ can be expressed in terms of $\Delta^2 u_{-1}$, $\Delta^3 u_{-1}$;

$\Delta^3 u_0$ in terms of $\Delta^3 u_{-1}$, $\Delta^4 u_{-2}$, $\Delta^5 u_{-2}$;

and so on, with the result:

$$\begin{aligned} u_x = u_0 + x\Delta u_0 + x_2\Delta^2 u_{-1} &+ (x+1)_3\Delta^3 u_{-1} \\ &+ (x+1)_4\Delta^4 u_{-2} + (x+2)_5\Delta^5 u_{-2} + \cdots \qquad (10) \end{aligned}$$

This interpolation formula is due to Gauss. It employs the odd differences just *below* the central line from u_0 and the even differences on the central line.

Another interpolation formula of Gauss may be derived in a similar manner; it employs the odd differences just *above* the central line from u_0 and the even differences on the central line, and is expressed as follows:

$$\begin{aligned} u_x = u_0 + x\Delta u_{-1} + (x+1)_2\Delta^2 u_{-1} &+ (x+1)_3\Delta^3 u_{-2} \\ &+ (x+2)_4\Delta^4 u_{-2} + (x+2)_5\Delta^5 u_{-3} + \cdots \qquad (11) \end{aligned}$$

The mean of (10) and (11) is

$$\begin{aligned} u_x = u_0 &+ x\left(\frac{\Delta u_{-1} + \Delta u_0}{2}\right) + \frac{x^2}{2}\Delta^2 u_{-1} \\ &+ (x+1)_3\left(\frac{\Delta^3 u_{-2} + \Delta^3 u_{-1}}{2}\right) + \frac{x}{4}(x+1)_3\Delta^4 u_{-2} \\ &+ (x+2)_5\left(\frac{\Delta^5 u_{-3} + \Delta^5 u_{-2}}{2}\right) + \frac{x}{6}(x+2)_5\Delta^6 u_{-3} + \cdots \qquad (12) \end{aligned}$$

This is the well-known central-difference formula, sometimes called Stirling's formula. Proceeding from the function u_0 in the diagonal difference table, it employs the mean of the odd differences above and below the central line and the even differences on the central line.[1]

Example 2. When Stirling's interpolation formula is applied to Table II, the area derived for (9) when $t = 1.22$ is

$$\begin{aligned} u_{.44} = .34134 &+ .44\left(\frac{.09185 + .14988}{2}\right) + .0968(-.05803) \\ &+ (-.05914)\left(\frac{.01024 - .01645}{2}\right) + (-.0065)(.02670) \\ = .34134 &+ .05318 - .00562 + .00018 - .00017 = .38891. \end{aligned}$$

$$\text{Error} = +.00014.$$

Bessel's interpolation formula. If u_x is obtained from Newton's formula by setting out from u_1 after inverting the columns of the diagonal difference table, the result is to express u_x in the diagonal running upward and to the right which sets out from u_1 and passes through Δu_0. The formula is

$$\begin{aligned} u_x = u_1 + (x-1)\Delta u_0 + x_2\Delta^2 u_{-1} &+ (x+1)_3\Delta^3 u_{-2} \\ &+ (x+2)_4\Delta^4 u_{-3} + (x+3)_5\Delta^5 u_{-4} + \cdots \qquad (13) \end{aligned}$$

[1] William Chauvenet, *Spherical and Practical Astronomy*, vol. 1, pp. 79–85.

and if the mean of this and Newton's formula is taken, Bessel's formula is obtained.[1]

$$u_x = u_0 + x\Delta u_0 + x_2\left(\frac{\Delta^2 u_{-1} + \Delta^2 u_0}{2}\right) + \frac{1}{3}x_2\left(x - \frac{1}{2}\right)\Delta^3 u_{-1}$$
$$+ (x+1)_4\left(\frac{\Delta^4 u_{-2} + \Delta^4 u_{-1}}{2}\right) + \frac{1}{5}(x+1)_4\left(x - \frac{1}{2}\right)\Delta^5 u_{-2} + \cdots \quad (14)$$

To obtain the terms in Bessel's formula from the diagonal difference table draw lines just below u_0 and above u_1 and take the odd differences inclosed between these lines and the mean of the even differences just above and below these lines.[2]

Example 3. Bessel's interpolation formula gives the area (9) under the normal curve of error when $t = 1.22$ as follows:

$$u_{.44} = .34134 + .44(,09185) + (-.1232)\left(\frac{-.04779 - .05803}{2}\right)$$
$$+ .00246(.01024) + .02306\left(\frac{.01003 + .02670}{2}\right)$$
$$= .34134 + .04041 + .00652 + .00003 + .00042 = .38872.$$
$$\text{Error} = .00005.$$

The reader is referred to the treatise of Rice[3] for a detailed explanation of difference tables and proofs of the interpolation formulas of Newton, Stirling, and Bessel. This book contains numerous applications of these formulas to problems in practical astronomy and will be found very helpful and suggestive.

Everett's interpolation formula. Another formula to which attention should be called was proposed by Everett.[4]

$$u_x = \xi u_0 + (\xi + 1)_3\Delta^2 u_{-1} + (\xi + 2)_5\Delta^4 u_{-2} + (\xi + 3)_7\Delta^6 u_{-3} + \cdots$$
$$+ x u_1 + (x+1)_3\Delta^2 u_0 + (x+2)_5\Delta^4 u_{-1} + (x+3)_7\Delta^6 u_{-2} + \cdots \quad (15)$$

where $\xi = 1 - x.$

In this formula only even differences occur and they lie along horizontal lines drawn through u_0 and u_1, respectively, in the diagonal difference table.

Since

$$\xi u_0 + x u_1 = u_0 + x\Delta u_0,$$

[1] Elias Loomis, *An Introduction to Practical Astronomy*, pp. 202–07.

[2] William Chauvenet, *loc. cit.*, pp. 86–87.

[3] Herbert L. Rice, *The Theory and Practice of Interpolation*, Chaps. I and II, pp. 1–95.

[4] J. D. Everett, "On a New Interpolation Formula," *Journal of the Institute of Actuaries*, vol. 35, pp. 452–58.

this formula may be written in the following convenient form for computation:

$$\begin{aligned} u_x = u_0 + x\Delta u_0 &+ (x+1)_3\Delta^2 u_0 \quad + (\xi+1)_3\Delta^2 u_{-1} \\ &+ (x+2)_5\Delta^4 u_{-1} + (\xi+2)_5\Delta^4 u_{-2} \\ &+ (x+3)_7\Delta^6 u_{-2} + (\xi+3)_7\Delta^6 u_{-3} + \cdots \end{aligned} \qquad (16)$$

Example 4. Everett's interpolation formula gives the area (9) under the normal curve of error when $t = 1.22$ as follows:

$$\begin{aligned} u_{.44} &= .34134 + .44(.09185) + (-.05914)(-.04779) \\ &\quad + (-.06406)(-.05803) + .01125(.01003) + .01181(.02670) \\ &= .34134 + .04041 + .00283 + .00372 + .00011 + .00032 = .38873. \\ &\text{Error} = -.00004. \end{aligned}$$

The choice of interpolation formulas. Where possible a central-difference formula should be used which involves the ordinates preceding and following the ordinate u_0 selected as the origin. No hard-and-fast rule can be laid down as to the choice between central-difference formulas, although it may be remarked that in practice Everett's and Bessel's formulas are usually preferred when the interpolated ordinate is for a value of x in the interval .25 to .75, particularly for values near .50. When $x = .50$, the coefficients of odd differences vanish in Bessel's formula.

Newton's formula should be used only when interpolated values are required near the beginning of a set of tabular ordinates. When the interpolated values desired are near the end of the set, Newton's formula may be applied to the ordinates inverted in order.

All these formulas express an ordinate at one point in terms of the ordinate at another point and the distance x between them and certain differences of the ordinates near the ordinate from which the interpolation sets out. The formulas have been illustrated with respect to the diagonal difference table, but it is a simple matter to pick out the corresponding terms in the horizontal difference table, and this form is suggested as more desirable to use in practice. For example, in Newton's formula, the coefficients are 1, x, x_2, x_3, and so on, and the differences are in the same row with the function from which we set out.

In the first formula of Gauss, the coefficients are 1, x, x_2, $(x+1)_3$, $(x+1)_4$, $(x+2)_5$, $\cdots$ with function and first difference in the same row, next two differences step up one row, next two step up another row, and so on. Simple rules for deriving the coefficients in most of these formulas are given by Chauvenet.[1]

Tables of values to five places of decimals of the coefficients up to

[1] William Chauvenet, *loc. cit.*, pp. 79–91.

and including fifth differences in Newton's, Stirling's, and Bessel's and sixth differences in Everett's formulas with the argument x varying by intervals of .01 from 0 to 1 are given by Glover.[1] Extensive tables to ten places of decimals up to and including sixth differences in Everett's formula with the argument varying by intervals of .001 from 0 to 1 are given by Thompson.[2]

SYSTEMATIC INTERPOLATION BY CONTINUOUS METHODS

Subdivision of intervals. The formulas already given are useful when only a few values are to be interpolated. When, however, a table is to be extended by the interpolation of many values, for example, nine new values between every pair of original values, it is better to adopt a continuous process. This is accomplished by expressing the leading differences of the subdivided minor intervals in terms of the leading differences of the original major intervals. The following example illustrates this process.

Example 5. The values of the Gamma Function are given at intervals of .1 from 1 to 1.5, both inclusive, and the continuous process is employed to interpolate nine values between 1.2 and 1.3.

The horizontal difference table for the Gamma Function and major differences is as follows:

TABLE IV

x	$\Gamma(x)$	$\Delta\Gamma(x)$	$\Delta^2\Gamma(x)$	$\Delta^3\Gamma(x)$	$\Delta^4\Gamma(x)$	$\Delta^5\Gamma(x)$
1.0	1.0000000	− .0216593	.0062411	− .0007251	.0001438	− .0000376
1.1	.9783407	− .0154182	.0055160	− .0005813	.0001062	
1.2	.9629225	− .0099022	.0049347	− .0004751		
1.3	.9530203	− .0049675	.0044596			
1.4	.9480528	− .0005079				
1.5	.9475449					

Subdivision into ten minor intervals by Newton's interpolation formula. The leading minor differences are then calculated by the following formula:[3]

$$\delta^k u_0 = a\Delta u_{-2} + b\Delta^2 u_{-2} + c\Delta^3 u_{-2} + d\Delta^4 u_{-2} + e\Delta^5 u_{-2} \qquad (17)$$

[1] James W. Glover, *Tables of Applied Mathematics in Finance, Insurance and Statistics*, pp. 412–19.

[2] A. J. Thompson, *Table of Coefficients of Everett's Central-Difference Interpolation Formula.* Tracts for Computers, Edited by Karl Pearson, No. V (1921).

[3] George King, *Textbook of the Institute of Actuaries*, Part II, second edition, pp. 441–53.

based on Newton's interpolation formula for subdividing the middle major interval into ten minor intervals. The coefficients are given in the following table:[1]

TABLE V

$\delta^k u$	a	b	c	d	e
δu_0	.1000	.1550	.0385	− .0086625	. .00329175
$\delta^2 u_0$		.0100	.0110	− .0002750	− .00024750
$\delta^3 u_0$			.0010	.0006500	− .00023750
$\delta^4 u_0$				.0001000	. .00002000
$\delta^5 u_0$					.00001000

These coefficients are obtained by using Newton's formula[2] (7) to interpolate nine equidistant ordinates between u_0 and u_1. The interpolated values are then differenced to express the leading minor differences of u_0 in terms of the leading major differences in u_0. Finally, employing (5), the leading major differences of u_0 are expressed in terms of the leading major differences of u_{-2}.

Having the value of the Gamma Function for the argument 1.2, namely, $u_0 = .9629225$, and the first five leading minor differences at this point, the nine interpolated values are easily calculated. The leading fifth minor difference vanishes and the leading fourth minor difference is equal to 1 in the eighth decimal place. Taking the fourth minor difference as constant and employing (5), the interpolated values are easily obtained by successive addition, as shown below. Decimal points are omitted in the calculation.

TABLE VI

$\delta^4\Gamma(x)$	$\delta^3\Gamma(x)$	$\delta^2\Gamma(x)$	$\delta\Gamma(x)$	$\Gamma(x)$	x
1	− 62	5441	− 122785	96292250	1.20
	− 61	5379	− 117344	96169465	1.21
	− 60	5318	− 111965	96052121	1.22
	− 59	5258	− 106647	95940156	1.23
	− 58	5199	− 101389	95833509	1.24
	− 57	5141	− 96190	95732120	1.25
	− 56	5084	− 91049	95635930	1.26
	− 55	5028	− 85965	95544881	1.27
		4973	− 80937	95458916	1.28
			− 75964	95377979	1.29
				95302015	1.30

[1] James W. Glover, *Tables of Applied Mathematics* (1923), p. 428.

[2] Alfred Henry, *Calculus and Probability* (1922), pp. 23–26.

The interpolated values of the Gamma Function obtained by this method are correct to six places of decimals.

Osculatory interpolation formulas. The interpolations in the interval 1.2 to 1.3 were derived from the leading minor differences of 1.2, which in turn were derived from the leading major differences of 1.0. These leading major differences were calculated from the six ordinates of the function from $x = 1.0$ to 1.5. In a similar manner the interpolations in the interval 1.3 to 1.4 would be based on the values of the six ordinates from $x = 1.1$ to 1.6; the interpolations in the interval 1.4 to 1.5, on the six ordinates from $x = 1.2$ to 1.7, and so on. The nine interpolations in *each interval* necessarily run smoothly because they lie on the same parabola, however, from interval to interval different parabolas are used and at their points of junction breaks in the continuity or smoothness of the interpolated values are likely to occur.

This difficulty is overcome by what is known as osculatory interpolation. The coefficients in the osculatory formulas are determined by the condition that at junction points the successive interpolation parabolas shall have a common ordinate, slope and (in case of fifth differences) osculating circle. This is effected by making the first and second derivatives of each successive pair of interpolation parabolas equal, respectively, at the common ordinate, that is, at the junction points. The third and fifth difference osculatory interpolation formulas are:

$$u_x = u_{-1} + (x+1)\Delta u_{-1} + (x+1)_2\Delta^2 u_{-1} + \frac{x^2(x-1)}{\underline{|2}}\Delta^3 u_{-1}, \quad (18)$$

$$u_x = u_{-2} + (x+2)\Delta u_{-2} + (x+2)_2\Delta^2 u_{-2} + (x+2)_3\Delta^3 u_{-2} + (x+2)_4\Delta^4 u_{-2} + \frac{x^3(x-1)(5x-7)}{\underline{|4}}\Delta^5 u_{-2}. \quad (19)$$

It will be observed that (18) is the same as Newton's formula in terms of u_{-1} and its leading differences except for the coefficient of the *third* difference; similarly (19) is the same as Newton's formula expressed in terms of u_{-2} and its leading differences except for the coefficient of the *fifth* difference.

Fifth difference osculatory interpolation. The following example illustrates the method of continuous interpolation by (19).

Example 6. The values of the area to the right of the origin under the normal curve of error are given by major intervals of 0.2 from $t = 0$ to 1.2 to effect fifth difference osculatory interpolation in minor intervals of .04 between the third and fourth ordinates and the fourth and fifth ordinates. The horizontal difference table is given in Table VII.

TABLE VII

t	x	u_x	Δu_x	$\Delta^2 u_x$	$\Delta^3 u_x$	$\Delta^4 u_x$	$\Delta^5 u_x$
.00	− 2	.000000	.079260	− .003098	− .002739	.000649	.000170
.20	− 1	.079260	.076162	− .005837	− .002090	.000819	.000035
.40	0	.155422	.070325	− .007927	− .001271	.000854	
.60	1	.225747	.062398	− .009198	− .000417		
.80	2	.288145	.053200	− .009615			
1.00	3	.341345	.043585				
1.20	4	.384930					

The leading major differences are used to compute the leading minor differences and with the latter the interpolated values are obtained by successive addition. The following table[1] gives the values of the coefficients of leading major differences in any row to compute the leading minor differences for the argument two rows below when six ordinates are given and the fifth difference osculatory interpolation formula is used for subdivision of the major interval into five minor intervals.

TABLE VIII

$\delta^k u$	a	b	c	d	e
δu	.2000	.3200	.0880	− .0176	.0016
$\delta^2 u$		.0400	.0480	.0016	.0048
$\delta^3 u$			.0080	.0064	− .0048
$\delta^4 u$				.0016	− .0032
$\delta^5 u$					.0080

When these coefficients are used with the leading major differences of u_{-2}, the leading minor differences of u_0 are obtained; when they are used with the leading major differences of u_{-1}, the leading minor differences of u_1 are obtained. Since u_0 and u_1 are known, the interpolated values in the intervals u_0 to u_1, and u_1 to u_2 are obtained by continuous addition as shown in Table IX.

The numbers in italics, *.2257474* and *.2881458* would have been reproduced exactly as .225747 and .288145, respectively, if all the figures in the computed leading minor differences had been retained; this is quite unnecessary, however, as seven places of decimals in these differences is sufficient to insure accuracy to five places in the interpolated values. Similar remarks apply to the italicized number *95302015* in Table VI.

[1] James W. Glover, *loc. cit.*, p. 428.

TABLE IX

$\delta^5 u$	$\delta^4 u$	$\delta^3 u$	$\delta^2 u$	δu	u_x	x	t
.0000014	.0000005	− .0000186	− .0002535	.0146085	.155422	.0	.40
	.0000019	− .0000181	− .0002721	.0143550	.1700305	.2	.44
		− .0000162	− .0002902	.0140829	.1843855	.4	.48
			− .0003064	.0137927	.1984684	.6	.52
				.0134863	.2122611	.8	.56
					.2257474	1.0	.60
.0000003	.0000012	− .0000116	− .0003323	.0131663	.225747	1.0	.60
	.0000015	− .0000104	− .0003439	.0128340	.2389133	1.2	.64
		− .0000089	− .0003543	.0124901	.2517473	1.4	.68
			− .0003632	.0121358	.2642374	1.6	.72
				.0117726	.2763732	1.8	.76
					.2881458	2.0	.80

References [1] on osculatory interpolation are given below for the reader who wishes to pursue the subject further.

Subdivision into halves.[2] The subdivision of an interval is sometimes performed by successively dividing the interval into halves. This process is called *interpolation into the middle.* When $x = \frac{1}{2}$ in Bessel's formula, the middle ordinate becomes

$$u_{\frac{1}{2}} = \frac{1}{2}(u_0 + u_1) - \frac{1}{8}\left(\frac{\Delta^2 u_0 + \Delta^2 u_{-1}}{2}\right) + \frac{3}{128}\left(\frac{\Delta^4 u_{-1} + \Delta^4 u_{-2}}{2}\right), \tag{20}$$

or in decimal form

$$u_{.5} = .5(u_0 + u_1) - .125\left(\frac{\Delta^2 u_0 + \Delta^2 u_{-1}}{2}\right) + .234375\left(\frac{\Delta^4 u_{-1} + \Delta^4 u_{-2}}{2}\right). \tag{21}$$

[1] Thomas Bond Sprague, "Explanation of a New Formula for Interpolation," *Journal of the Institute of Actuaries*, vol. 22, pp. 270–85.

Johannes Karup, "On a New Mechanical Method of Graduation," *Transactions of the Second International Actuarial Congress*, pp. 78–109.

George King, "On the Construction of Mortality Tables from Census Returns and Records of Deaths," *Journal of the Institute of Actuaries*, vol. 42, pp. 225–77.

George King, "On a New Method of Constructing and Graduating Mortality and Other Tables," *Journal of the Institute of Actuaries*, vol. 43, pp. 109–84.

James Buchanan, "Osculatory Interpolation by Central Differences: with an Application to Life Table Construction," *Journal of the Institute of Actuaries*, vol. 42, pp. 369–94.

George J. Lidstone, "Alternative Demonstration of the Formula for Osculatory Interpolation," *Journal of the Institute of Actuaries*, vol. 42, pp. 394–400.

James W. Glover, "Derivation of the United States Mortality Table by Osculatory Interpolation," *Quarterly Publication of American Statistical Association*, vol. 12, pp. 85–109.

United States Life Tables, 1890, 1901, 1910, and 1901–10, pp. 344–48 and 372–88.

[2] William Chauvenet, *loc. cit.*, pp. 87–88. Herbert L. Rice, *loc. cit.*, pp. 78–96.

Since the coefficients of differences of odd order in Bessel's formula vanish when $x = \frac{1}{2}$, the first two terms of the above formula give the middle ordinate correct to third differences and the first three terms correct to fifth differences.

Formula (20) is frequently employed to halve intervals and in many textbooks is put in the form of a rule. The rule may be stated as follows for the horizontal difference table:

From the mean of the two given functions subtract $\frac{1}{8}$ *of the mean of the second differences in the same row and the row above.* This result is correct to third differences inclusive. To obtain the value correct to fifth differences inclusive, *add to the above result* $\frac{3}{128}$ *of the mean of the fourth differences in the first and second rows above.*

By n successive applications of this formula the original intervals are subdivided into 2^n intervals.

Example 7. Subdivide the original intervals of .16 into intervals of .04 between .40 and .80 by two applications of the formula to the integral of the normal curve of error given in (9).

TABLE X

t	u_x	Δu_x	$\Delta^2 u_x$	t	u_x	Δu_x	$\Delta^2 u_x$
.16	.063560	.061956	− .003086	.32	.125516	.029899	− .000928
.32	.125516	.058870	− .004342	.40	.155415	.028971	− .001105
.40	*.155415*			*.44*	*.170028*		
.48	.184386	.054528	− .005297	.48	.184386	.027866	− .001204
.56	*.212252*			*.52*	*.198463*		
.64	.238914	.049231	− .005904	.56	.212252	.026662	− .001346
.72	*.264230*			*.60*	*.225742*		
.80	.288145	.043327	− .006156	.64	.238914	.025316	− .001401
.88	*.310563*			*.68*	*.251744*		
.96	.331472	.037171		.72	.264230	.023915	− .001497
				.76	*.276369*		
1.12	.368643			.80	.288145	.022418	
				.88	.310563		

The interpolation on the left in Table X is made first, the italicized numbers being obtained. These numbers, together with the original numbers from .32 to .88 are then differenced again on the right and a second interpolation gives the additional italicized numbers. Only second differences are used in each application, but the results are good to third differences inclusive because the coefficient of the latter vanishes in Bessel's formula when $x = \frac{1}{2}$. Although only second differences and six decimal places were used, the final results are correct within a unit in the fifth decimal place.

Errors in tables. The detection of accidental errors in tables is facilitated by constructing the difference columns. This is shown most easily by the following table containing the error e in the value of u_3.

TABLE XI[1]

u_0			$\Delta^3 u_0 + e$
u_1		$\Delta^2 u_1 + e$	$\Delta^3 u_1 - 3\,e$
u_2	$\Delta u_2 + e$	$\Delta^2 u_2 - 2\,e$	$\Delta^3 u_2 + 3\,e$
$u_3 + e$	$\Delta u_3 - e$	$\Delta^2 u_3 + e$	$\Delta^3 u_3 - e$
u_4	Δu_4	$\Delta^2 u_4$	$\Delta^3 u_4$

It will be observed that the error, e, is multiplied in the difference column of a given order by the binomial coefficients of the same order with alternating signs. Accordingly, when the correct differences approach a constant value the succeeding difference column will show up a succession of alternating differences closely following in value the product of the error and the binomial coefficients of the order of differences in that column.

Ordinates not equidistant. All the interpolation formulas given so far are based on the assumption of equidistant ordinates. If the ordinates are not equidistant, it is still possible to determine a rational integral algebraic function which will pass through the ends of these ordinates; the function so determined defines the interpolated ordinates for values of the independent variable. Let the ordinates be u_a, u_b, u_c, $\cdots$, and let

$$u_x = A + Bx + Cx^2 + Dx^3 + \cdots$$

Then when a, b, c are substituted in the above equation, a set of linear equations is obtained which gives the required values of the coefficients $A, B, C, \cdots$

This interpolation formula was given by Lagrange in the form

$$\begin{aligned} u_x = {} & u_a \frac{(x-b)(x-c)\cdots(x-n)}{(a-b)(a-c)\cdots(a-n)} \\ & + u_b \frac{(x-a)(x-c)\cdots(x-n)}{(b-a)(b-c)\cdots(b-n)} \\ & + \cdots \end{aligned} \tag{22}$$

and it is usually called **Lagrange's formula.**

An inspection shows that the right-hand member is a rational integral algebraic function of x of degree one less than the number of ordinates

[1] Herbert L. Rice, *loc. cit.*, pp. 9–15.

and it is also evident that it equals u_a when x equals a, u_b when x equals b, and so on.

This formula may be used to supply the unknown value of the argument when the function is taken as abscissa, for example, when a logarithm is given to find the anti-logarithm. Here the abscissa x is the logarithm and the ordinate y the anti-logarithm. In short, any functional relation between two variables may be assumed to be determined by a parabolic curve for a small range and Lagrange's formula used to determine the value of the unknown ordinate.[1]

Example 8. Given the following table, to find the unknown ordinate corresponding to the abscissa .86614.

Table XII

t	$x = \frac{2}{\sqrt{\pi}} \int_0^t e^{-t^2} dt$	$u_x = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-t^2/2}\, dt$
1.00	.84270	.34134
1.04	.85865	.35083
1.06	.86614	
1.13	.88997	.37076

Applying Lagrange's formula, with

$$a = .84270 \qquad u_a = .34134$$
$$b = .85865 \qquad u_b = .35083$$
$$c = .88997 \qquad u_c = .37076$$
$$x = .86614$$

$$u_x = .35544 \qquad \text{Correct value } u_x = .35543$$

Numerous writers have attempted to simplify the notation of the calculus of differences and present the various types of central and noncentral formulas in a systematic and logical order. References[2] are

[1] Alfred Henry, *loc. cit.*, pp. 51–53.

[2] W. F. Sheppard, "Central-Difference Formulæ," *Proceedings, London Mathematical Society*, vol. 31, pp. 449–88.

W. F. Sheppard, "Central-Difference Interpolation Formulæ," *Journal of the Institute of Actuaries*, vol. 50, pp. 85–89.

Robert Henderson, "A Practical Interpolation Formula with a Theoretical Introduction," *Transactions of the Actuarial Society of America*, vol. 9, pp. 211–24.

S. A. Joffe, "Interpolation-Formulæ and Central-Difference Notation," *Transactions of the Actuarial Society of America*, vol. 18, pp. 72–98.

S. A. Joffe, "Parallel Proofs of Everett's, Gauss's, and Newton's Central-Difference Interpolation-Formulæ," *Transactions of the Actuarial Society of America*, vol. 20, pp. 423–29.

given to some of these papers and there will be found therein further references which practically cover this field.

Rational integral algebraic functions. The nth difference of a rational integral algebraic function of the nth degree is constant and all higher differences vanish. Let

$$u_x = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n.$$

Then

$$\Delta u_x = na_nx^{n-1}$$

plus terms of lower degree and the truth of the above statement follows at once by successive differencing. The constant nth difference is evidently equal to $a_n\underline{|n}$, or $\omega^n\, a_n\underline{|n}$ when the interval from u_0 to u_1 contains ω units of the argument of the function. Since

$$\Delta u_x = u_{x+1} - u_x$$

and

$$\Delta^2 u_x = \Delta u_{x+1} - \Delta u_x = u_{x+2} - 2\, u_{x+1} + u_x$$

and

$$\Delta^{k+1} u_x = \Delta \cdot \Delta^k u_x,$$

it follows easily by mathematical induction that

$$\Delta^n u_x = u_{x+n} - nu_{x+n-1} + n_2u_{x+n-2} + \cdots + (-1)^n u_x.$$

If $x = 0$, we have

$$\Delta^n u^0 = u_n - n_1u_{n-1} + n_2u_{n-2} + \cdots + (-1)^n u_0, \qquad (23)$$

which expresses the nth difference of u^0 in terms of u_0 and the n functions following u_0.

This formula [1] is useful in interpolating one or more missing terms in a series of equidistant terms. For example, if one term is missing, a curve of degree $n - 1$ can be passed through the n terms which are given. The nth differences of this function vanish, hence the missing or unknown term is found by solving this equation for that term.

If two terms are missing, a curve of degree $n - 2$ can be passed through the $n - 1$ terms which are given. The $(n - 1)$th differences of this function vanish, hence two linear equations can be written as follows:

$$\left.\begin{aligned}\Delta^{n-1}u_0 &= u_{n-1} - (n-1)_1u_{n-2} + (n-2)_2u_{n-3} \cdots + (-1)^{n-1}u_0 = 0,\\ \Delta^{n-1}u_1 &= u_n - (n-1)_1u_{n-1} + (n-2)_2u_{n-2} \cdots + (-1)^n u_1 = 0,\end{aligned}\right\} \qquad (24)$$

whose solution will give the two missing terms. The extension of this principle is obvious.

[1] Burn and Brown, *Elements of Finite Differences*, pp. 23–24.

DERIVATIVES IN TERMS OF DIFFERENCES

Derivatives by Newton's formula. It is frequently required to determine the derivative of a function in terms of its finite differences. By differentiating u_x with respect to x as expressed by Newton's and other interpolation formulas given, the first and higher derivatives can be obtained in terms of the differences of u_x. Applying this process to Newton's formula (7), the following equations are obtained expressing the values of the several derivatives in terms of the differences of u_0 and the abscissa x. The number of units in the interval u_0 to u_1 is denoted by ω.

$$\omega u'_x = \omega \frac{du_x}{dx} = \Delta u_0 + \frac{dx_2}{dx}\Delta^2 u_0 + \frac{dx_3}{dx}\Delta^3 u_0 + \frac{dx_4}{dx}\Delta^4 u_0 + \frac{dx_5}{dx}\Delta^5 u_0 + \cdots$$

Hence,

$$\begin{aligned}\omega u'_x = \Delta u_0 &+ (x - \tfrac{1}{2})\Delta^2 u_0 + (\tfrac{1}{2}x^2 - x + \tfrac{1}{3})\Delta^3 u_0 \\ &+ (\tfrac{1}{6}x^3 - \tfrac{3}{4}x^2 + \tfrac{11}{12}x - \tfrac{1}{4})\Delta^4 u_0 \\ &+ (\tfrac{1}{24}x^4 - \tfrac{1}{3}x^3 + \tfrac{7}{8}x^2 - \tfrac{5}{6}x + \tfrac{1}{5})\Delta^5 u_0 + \cdots\end{aligned} \qquad (25)$$

Similarly,

$$\begin{aligned}\omega^2 u''_x = \Delta^2 u_0 &+ (x - 1)\Delta^3 u_0 + (\tfrac{1}{2}x^2 - \tfrac{3}{2}x + \tfrac{11}{12})\Delta^4 u_0 \\ &+ (\tfrac{1}{6}x^3 - x^2 + \tfrac{7}{4}x - \tfrac{5}{6})\Delta^5 u_0 + \cdots\end{aligned} \qquad (26)$$

$$\omega^3 u'''_x = \Delta^3 u_0 + (x - \tfrac{3}{2})\Delta^4 u_0 + (\tfrac{1}{2}x^2 - 2x + \tfrac{7}{4})\Delta^5 u_0 + \cdots \qquad (27)$$

Example 9. Find the first and second derivatives of the function (9) when $t = 1.22$.

Direct substitution in (25) and (26) of the differences in Table 2 with $x = .44$ and $\omega = \frac{1}{2}$ gives for $t = 1.22$.

$$u'_{.44} = .18884 \qquad u''_{.44} = -.24284$$

The values of the derivatives at the ordinate u_0 are found by setting $x = 0$ in (25), (26), (27), and the resulting formulas are

$$\omega u'_0 = \Delta u_0 - \tfrac{1}{2}\Delta^2 u_0 + \tfrac{1}{3}\Delta^3 u_0 - \tfrac{1}{4}\Delta^4 u_0 + \tfrac{1}{5}\Delta^5 u_0 + \cdots \qquad (28)$$

$$\omega^2 u''_0 = \Delta^2 u_0 - \Delta^3 u_0 + \tfrac{11}{12}\Delta^4 u_0 - \tfrac{5}{6}\Delta^5 u_0 + \cdots \qquad (29)$$

$$\omega^3 u'''_0 = \Delta^3 u_0 - \tfrac{3}{2}\Delta^4 u_0 + \tfrac{7}{4}\Delta^5 u_0 + \cdots \qquad (30)$$

Example 10. Find the first and second derivatives at $t = 1$ in the function defined by (9).

This value of t corresponds to $x = 0$ in Table II and substitution of the differences in the u_0 row in (28) and (29) gives

$$u'_0 = .24724 \qquad u''_0 = -.28852$$

Derivatives by Stirling's formula. When (12) is employed, the following formulas are obtained expressing the derivatives in terms of the differences.

$$\omega u'_x = \frac{\Delta u_{-1} + \Delta u_0}{2} + x\Delta^2 u_{-1} + (\tfrac{1}{2}x^2 - \tfrac{1}{6})\left(\frac{\Delta^3 u_{-2} + \Delta^3 u_{-1}}{2}\right)$$
$$+ (\tfrac{1}{6}x^3 - \tfrac{1}{12}x)\Delta^4 u_{-2} + (\tfrac{1}{24}x^4 - \tfrac{1}{8}x^2 + \tfrac{1}{30})\left(\frac{\Delta^5 u_{-3} + \Delta^5 u_{-2}}{2}\right) + \cdots \quad (31)$$

$$\omega^2 u''_x = \Delta^2 u_{-1} + x\left(\frac{\Delta^3 u_{-2} + \Delta^3 u_{-1}}{2}\right) + (\tfrac{1}{2}x^2 - \tfrac{1}{12})\,\Delta^4 u_{-2}$$
$$+ (\tfrac{1}{6}x^3 - \tfrac{1}{4}x)\left(\frac{\Delta^5 u_{-3} + \Delta^5 u_{-2}}{2}\right) + \cdots \quad (32)$$

$$\omega^3 u'''_x = \frac{\Delta^3 u_{-2} + \Delta^3 u_{-1}}{2} + x\Delta^4 u_{-2} + (\tfrac{1}{2}x^2 - \tfrac{1}{4})\left(\frac{\Delta^5 u_{-3} + \Delta^5 u_{-2}}{2}\right) + \cdots \quad (33)$$

Example 11. The first and second derivatives of (9) for $t = 1.22$ by (31) and (32) are

$u'_{.44} = .18992 \qquad u''_{.44} = -.23616$

When $x = 0$, Stirling's formulas for the first three derivatives simplify to the following:

$$\omega u'_0 = \frac{\Delta u_{-1} + \Delta u_0}{2} - \tfrac{1}{6}\left(\frac{\Delta^3 u_{-1} + \Delta^3 u_{-2}}{2}\right) + \tfrac{1}{30}\left(\frac{\Delta^5 u_{-2} + \Delta^5 u_{-3}}{2}\right) + \cdots \quad (34)$$

$$\omega^2 u''_0 = \Delta^2 u_{-1} - \tfrac{1}{12}\Delta^4 u_{-2} + \cdots \quad (35)$$

$$\omega^3 u'''_0 = \frac{\Delta^3 u_{-1} + \Delta^3 u_{-1}}{2} - \tfrac{1}{4}\left(\frac{\Delta^5 u_{-3} + \Delta^5 u_{-2}}{2}\right) + \cdots \quad (36)$$

Example 12. The first and second derivatives of (9) for $t = 1$ by (34) and (35) are

$u'_0 = .24278 \qquad u''_0 = -.24104$

Derivatives by Bessel's formula. When (14) is differentiated, the following formulas are obtained for the derivatives at the point x.

$$\omega u'_x = \Delta u_0 + (x - \tfrac{1}{2})\left(\frac{\Delta^2 u_{-1} + \Delta^2 u_0}{2}\right) + (\tfrac{1}{2}x^2 - \tfrac{1}{2}x + \tfrac{1}{12})\Delta^3 u_{-1}$$
$$+ (\tfrac{1}{6}x^3 - \tfrac{1}{4}x^2 - \tfrac{1}{12}x + \tfrac{1}{12})\left(\frac{\Delta^4 u_{-2} + \Delta^4 u_{-1}}{2}\right)$$
$$+ (\tfrac{1}{24}x^4 - \tfrac{1}{12}x^3 + \tfrac{1}{24}x - \tfrac{1}{120})\Delta^5 u_{-2} + \cdots \quad (37)$$

$$\omega^2 u''_x = \frac{\Delta^2 u_{-1} + \Delta^2 u_0}{2} + (x - \tfrac{1}{2})\Delta^3 u_{-1} + (\tfrac{1}{2}x^2 - \tfrac{1}{2}x - \tfrac{1}{12})\left(\frac{\Delta^4 u_{-2} + \Delta^4 u_{-1}}{2}\right)$$
$$+ (\tfrac{1}{6}x^3 - \tfrac{1}{4}x^2 + \tfrac{1}{24})\Delta^5 u_{-2} + \cdots \quad (38)$$

$$\omega^3 u'''_x = \Delta^3 u_{-1} + (x - \tfrac{1}{2})\left(\frac{\Delta^4 u_{-2} + \Delta^4 u_{-1}}{2}\right) + (\tfrac{1}{2}x^2 - \tfrac{1}{2}x)\Delta^5 u_{-2} + \cdots \quad (39)$$

Example 13. The first and second derivatives of (9) for $t = 1.22$ by (37) and (38) are

$u'_{.44} = .18968 \qquad u'_{.44} = -.22924$

When $x = 0$, Bessel's formulas for the first three derivatives may be written as follows:

$$\begin{aligned}\omega u_0' = \Delta u_0 - \tfrac{1}{2}\left(\frac{\Delta^2 u_{-1} + \Delta^2 u_0}{2}\right) + \tfrac{1}{12}\,\Delta^3 u_{-1} \\ + \tfrac{1}{12}\left(\frac{\Delta^4 u_{-2} + \Delta^4 u_{-1}}{2}\right) - \tfrac{1}{120}\,\Delta^5 u_{-2} + \cdots\end{aligned} \tag{40}$$

$$\omega^2 u_0'' = \frac{\Delta^2 u_{-1} + \Delta^2 u_0}{2} - \tfrac{1}{2}\,\Delta^3 u_{-1} - \tfrac{1}{12}\left(\frac{\Delta^4 u_{-2} + \Delta^4 u_{-1}}{2}\right) + \tfrac{1}{24}\,\Delta^5 u_{-2} + \cdots \tag{41}$$

$$\omega^3 u_0''' = \Delta^3 u_{-1} - \tfrac{1}{2}\left(\frac{\Delta^4 u_{-2} + \Delta^4 u_{-1}}{2}\right) + \cdots \tag{42}$$

Example 14. The first and second derivatives of (9) for $t = 1.22$ by (40) and (41) are

$$u_0' = .24138 \qquad\qquad u_0'' = -.23824$$

Derivatives by Everett's formula. When (14) is differentiated formulas for the derivatives at the point x are obtained as follows, where $\xi = 1 - x$:

$$\begin{aligned}\omega u_x' = {} & \Delta u_0 + (\tfrac{1}{2}\,x^2 - \tfrac{1}{6})\Delta^2 u_0 - (\tfrac{1}{2}\,\xi^2 - \tfrac{1}{6})\Delta^2 u_{-1} \\ & + (\tfrac{1}{24}\,x^4 - \tfrac{1}{8}\,x^3 + \tfrac{1}{30})\Delta^4 u_{-1} - (\tfrac{1}{24}\,\xi^4 - \tfrac{1}{8}\,\xi^2 + \tfrac{1}{30})\Delta^4 u_{-2} \\ & + (\tfrac{1}{720}\,x^6 - \tfrac{1}{72}\,x^4 + \tfrac{7}{240}\,x^2 - \tfrac{1}{140})\Delta^6 u_{-2} \\ & - (\tfrac{1}{720}\,\xi^6 - \tfrac{1}{72}\,\xi^4 + \tfrac{7}{240}\,\xi^2 - \tfrac{1}{140})\Delta^6 u_{-3}.\end{aligned} \tag{43}$$

$$\begin{aligned}\omega^2 u_x'' = {} & x\Delta^2 u_0 + \xi\Delta^2 u_{-1} \\ & + (\tfrac{1}{6}\,x^3 - \tfrac{1}{4}\,x)\Delta^4 u_{-1} + (\tfrac{1}{6}\,\xi^3 - \tfrac{1}{4}\,\xi)\Delta^4 u_{-2} \\ & + (\tfrac{1}{120}\,x^5 - \tfrac{1}{18}\,x^3 + \tfrac{7}{120}\,x)\Delta^6 u_{-2} + (\tfrac{1}{120}\,\xi^5 - \tfrac{1}{18}\,\xi^3 + \tfrac{7}{120}\,\xi)\Delta^6 u_{-3}.\end{aligned} \tag{44}$$

$$\begin{aligned}\omega^3 u_x''' = {} & \Delta^2 u_0 - \Delta^2 u_{-1} \\ & + (\tfrac{1}{2}\,x^2 - \tfrac{1}{4})\Delta^4 u_{-1} - (\tfrac{1}{2}\,\xi^2 - \tfrac{1}{4})\Delta^4 u_{-2} \\ & + (\tfrac{1}{24}\,x^4 - \tfrac{1}{6}\,x^2 + \tfrac{7}{120})\Delta^6 u_{-2} - (\tfrac{1}{24}\,\xi^4 - \tfrac{1}{6}\,\xi^2 + \tfrac{7}{120})\Delta^6 u_{-3}.\end{aligned} \tag{45}$$

Example 15. The first and second derivatives of (9) for $t = 1.22$ by (43) and (44) are

$$u'_{.44} = .18956 \qquad\qquad u''_{.44} = -.22980$$

When $x = 0$, Everett's formulas for the first three derivatives take the following form:

$$\begin{aligned}\omega u_0' = {} & \Delta u_0 - \tfrac{1}{6}\,\Delta^2 u_0 - \tfrac{1}{3}\,\Delta^2 u_{-1} \\ & + \tfrac{1}{30}\,\Delta^4 u_{-1} + \tfrac{1}{20}\,\Delta^4 u_{-2} \\ & - \tfrac{1}{140}\,\Delta^6 u_{-2} - \tfrac{1}{105}\,\Delta^6 u_{-3}.\end{aligned} \tag{46}$$

$$\omega^2 u_0'' = \Delta^2 u_{-1} - \tfrac{1}{12}\,\Delta^4 u_{-2} + \tfrac{7}{120}\,\Delta^6 u_{-3}. \tag{47}$$

$$\begin{aligned}\omega^3 u''' = {} & \Delta^2 u_0 - \Delta^2 u_{-1} \\ & - \tfrac{1}{4}\,\Delta^4 u_{-1} - \tfrac{1}{4}\,\Delta^4 u_{-2} \\ & + \tfrac{7}{120}\,\Delta^6 u_{-2} - \tfrac{29}{360}\,\Delta^6 u_{-3}.\end{aligned} \tag{48}$$

Example 16. The first and second derivatives of (9) for $t = 1$ by (46) and (47) are:

$$u_0' = .24166 \qquad u_0'' = -.24104$$

In examples 9 to 16 the first and second derivatives have been computed to fourth differences inclusive by employing four different interpolation formulas. The results are brought together for comparison in tabular form.

Table XIII

Formula	u_0'	Error	u_0''	Error	$u_{.44}'$	Error	$u_{.44}''$	Error
True Value	.24197	0	− .24197	0	.18954	0	− .23124	0
Everett's	.24166	− 31	− .24104	93	.18956	2	− .22980	144
Bessel's	.24138	− 59	− .23824	373	.18968	14	− .22924	200
Stirling's	.24278	81	− .24104	93	.18992	38	− .23616	− 492
Newton's	.24724	527	− .28852	− 4655	.18884	− 70	− .24284	− 1160

It will be observed that central-difference formulas give better results than Newton's and that the choice as between the former is in the order Everett, Bessel, Stirling. While this will not always be true, it may be regarded as confirming the remarks made on page 40 concerning choice of interpolation formulas.

A very full treatment of derivatives of tabular functions in terms of differences will be found in Rice.[1]

SUMMATION FORMULAS

Finite integrals in terms of differences. If the first column to the left of the column containing the function u_x is expanded by Newton's formula, the following result is obtained:

$$\Delta^{-1}u_x = \Delta^{-1}u_0 + x_1u_0 + x_2\Delta u_0 + x_3\Delta^2u_0 + \cdots$$

Taking into account the principle involved in statement (*e*) on page 35, we are led easily to the following expression for the sum of the terms in the u_x column from $x = 0$ to $x = 1$, inclusive,

$$\begin{aligned}\Delta^{-1}u_x - \Delta^{-1}u_0 &= \Delta^{-1}u_x\Big|_0^x = u_0 + u_1 + \cdots + u_{x-1} \\ &= \sum_0^{x-1}u_x = xu_0 + x_2\Delta u_0 + x_3\Delta^2u_0 + x_4\Delta^3u_0 + \cdots \qquad (49)\end{aligned}$$

Similarly, the sum of n terms beginning with u_a can be expressed as follows:

$$\begin{aligned}\sum_a^{a+n-1}u_x &= u_a + u_{a+1} + \cdots + u_{a+n-1} \\ &= nu_a + n_2\Delta u_a + n_3\Delta^2u_a + n_4\Delta^3u_a + \cdots \qquad (50)\end{aligned}$$

[1] Herbert L. Rice, *loc. cit.*, Chap. III, pp. 96–129.

In this manner the finite integral is expressed in terms of the finite differences. It is most useful when u_x is a rational integral algebraic function, for then the differences eventually vanish. The following example will illustrate an application of the formula.

Example 17. Compute a table of the first hundred values of the coefficient of $\Delta^4 u_0$ in (25),

$$x'_4 = \tfrac{1}{6}x^3 - \tfrac{3}{4}x^2 + \tfrac{11}{12}x - \tfrac{1}{4},$$

when the argument x proceeds by intervals of .01 from 0 to .99 inclusive, and check the results by (50).

The function x'_4 is first computed for the four values .00, .01, .02, .03 and the first three leading differences of u_0 formed as in the following table.

TABLE XIV

$\Delta^3 x'_4$	$\Delta^2 x'_4$	$\Delta x'_4$	$u_x = x'_4$	x
.000001	− .000149	.00909183	− .25000000	.00
	− .000148	.00894283	− .24090816	.01
		.00879483	− .23196533	.02
			− .22317050	.03

Since x'_4 is of the third degree, $\Delta^3 x'_4 = .000001$ is constant and x'_4 can be computed by intervals of .01 to $x = .99$ by successive additions. If these computed values are added at any stage in the work, the sum should check with the result obtained by substitution in (50). For example, the sum of the first hundred terms, setting $a = 0$ and $n = 100$, should be

$$\begin{aligned}\sum_0^{99} x'_4 &= 100\,u_0 + 100_2\Delta u_0 + 100_3\Delta^2 u_0 + 100_4\Delta^3 u_0\\ &= 100(-\,.25) + 4950(.00909183)\\ &\quad + 161700(-\,.000149) + 3921225(.000001)\\ &= -\,.1675.\end{aligned}$$

When x'_4 is computed by successive additions, a simple check on the work up to any point could be obtained of course by an independent computation of the function. If, however, the table had been constructed by computing the function independently for each individual value of the argument x, or if it were not known how the table had been computed, the check by (50) would be simple and effective except for compensated errors.

Lubbock's formula. The formula proposed by Lubbock[1] expresses a relation between the sum of a series of original and interpolated or-

[1] J. W. Lubbock, "On the Comparison of Various Tables of Annuities," *Transactions of the Cambridge Philosophical Society*, vol. 3, pp. 321–41. Reprinted in *Journal of the Institute of Actuaries*, vol. 5, pp. 277–92.

dinates, and the differences of the original ordinates between which the interpolations are made. It is written as follows:

$$
\begin{aligned}
\sum_{t=0}^{mn-1} u_{\frac{t}{n}} &= u_0 + u_{\frac{1}{n}} + u_{\frac{2}{n}} + \cdots + u_1 + u_{1+\frac{1}{n}} + \cdots + u_{m-\frac{1}{n}} \\
&= n(u_0 + u_1 + \cdots + u_{m-1}) \\
&+ \frac{n-1}{2}(u_m - u_0) - \frac{n^2-1}{12\,n}(\Delta u_m - \Delta u_0) + \frac{n^2-1}{24\,n}(\Delta^2 u_m - \Delta^2 u_0) \\
&- \frac{(n^2-1)(19\,n^2-1)}{720\,n^3}(\Delta^3 u_m - \Delta^3 u_0) + \frac{(n^2-1)(9\,n^2-1)}{480\,n^3}(\Delta^4 u_m - \Delta^4 u_0) \\
&- \frac{(n-1)(863\,n^4 - 145\,n^2 + 2)}{60480\,n^5}(\Delta^5 u_m - \Delta^5 u_0) \\
&+ \frac{(n^2-1)(275\,n^4 - 61\,n^2 + 2)}{24192\,n^5}(\Delta^6 u_m - \Delta^6 u_0) + \cdots \qquad (51)
\end{aligned}
$$

The original ordinates which are given, together with the differences of the first and last ordinates, are in the right-hand member. It will be observed that $n - 1$ ordinates are interpolated between each pair of the original ordinates from the ordinate u_0 to the ordinate u_m. If the function to be summed vanishes for the ordinates beginning with u_m, then u_m and all its differences vanish and the formula reduces to the following:

$$
\begin{aligned}
\sum_{t=0}^{mn-1} u_{\frac{t}{n}} &= u_0 + u_{\frac{1}{n}} + u_{\frac{2}{n}} + \cdots + u_1 + u_{1+\frac{1}{n}} + \cdots + u_{m-\frac{1}{n}} \\
&= n(u_0 + u_1 + u_2 + \cdots + u_{m-1}) - \frac{n-1}{2}u_0 + \frac{n^2-1}{12\,n}\Delta u_0 \\
&- \frac{n^2-1}{24\,n}\Delta^2 u_0 + \frac{(n^2-1)(19\,n^2-1)}{720\,n^3}\Delta^3 u_0 - \frac{(n^2-1)(9\,n^2-1)}{480\,n^3}\Delta^4 u_0 \\
&+ \frac{(n^2-1)(863\,n^4 - 145\,n^2 + 2)}{60480\,n^5}\Delta^5 u_0 \\
&- \frac{(n^2-1)(275\,n^4 - 61\,n^2 + 2)}{24192\,n^5}\Delta^6 u_0 + \cdots \qquad (52)
\end{aligned}
$$

Formula (52) is also used when u_m approaches zero asymptotically because then u_m and its differences become small and may be neglected.

Lubbock's formula is very useful in practical calculations where it is required to insert a given number of ordinates between every pair of original ordinates and obtain the sum of the original and interpolated ordinates.

Example 18. Given the value of the reciprocal of every fifth integer from 51 to 131, to find the sum of the reciprocals of all integers from 51 to 100, both inclusive.

Table XV

	n	$1/n$	Δ	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6
u_0	51	.019608	− .001751	.000287	− .000064	.000015	.000000	− .000006
u_1	56	.017857	− .001464	.000223	− .000049	.000015	− .000006	
u_2	61	.016393	− .001241	.000174	− .000034	.000009		
u_3	66	.015152	− .001067	.000140	− .000025			
u_4	71	.014085	− .000927	.000115				
u_5	76	.013158	− .000812					
u_6	81	.012346						
u_7	86	.011628						
u_8	91	.010989						
u_9	96	.010417						
u_{10}	101	.009901	− .000467	.000042	− .000005	− .000001	.000006	− .000016
u_{11}	106	.009434	− .000425	.000037	− .000006	.000005	− .000010	
u_{12}	111	.009009	− .000388	.000031	− .000001	− .000005		
u_{13}	116	.008621	− .000357	.000030	− .000006			
u_{14}	121	.008264	− .000327	.000024				
u_{15}	126	.007937	− .000303					
u_{16}	131	.007634						

Substituting in formula (51), the sum of the reciprocals of the integers from 51 to 100 is found to be .688172, which is correct to six places of decimals.

Values of the coefficients to eight places of decimals in Lubbock's formula for the first six differences from $n = 2$ to $n = 12$ are given by Glover.[1]

Woolhouse's formula. The Euler-Maclaurin expansion expresses a relation between the sum of the initial, terminal, and interpolated ordinates, the area under the curve as determined by the definite integral and the successive derivatives at the initial and terminal ordinates of the area in question. It may be written as follows:

$$\begin{aligned}\Sigma^{(m)}u_x &= \frac{1}{m}\left(u_0 + u_{\frac{1}{m}} + u_{\frac{2}{m}} + \cdots + u_{\omega - \frac{1}{m}} + u_\omega\right)\\ &= \int_0^\omega u_x\,dx + \frac{1}{2\,m}(u_0 + u_\omega) - \frac{1}{12\,m^2}\left(\frac{du_0}{dx} - \frac{du_\omega}{dx}\right)\\ &\quad + \frac{1}{720\,m^4}\left(\frac{d^3u_0}{dx^3} - \frac{d^3u_\omega}{dx^3}\right) - \frac{1}{30240\,m^6}\left(\frac{d^5u_0}{dx^5} - \frac{d^5u_\omega}{dx^5}\right) + \cdots \qquad (53)\end{aligned}$$

The initial ordinate is u_0, the terminal is u_ω, and the interpolated ordinates are at intervals of $\frac{1}{m}$ between u_0 and u_ω. The left-hand member of (53) is the sum of all these ordinates divided by m.

[1] James W. Glover, *Tables of Applied Mathematics*, p. 430.

If the extreme ordinate u_ω and the derivatives of u_x at $x = \omega$ vanish, (53) simplifies to the form:

$$\Sigma^{(m)}u_x = \int_0^\omega u_x dx + \frac{1}{2\,m}u_0 - \frac{1}{12\,m^2}\frac{du_0}{dx} + \frac{1}{720\,m^4}\frac{d^3u_0}{dx^3} - \frac{1}{30240\,m^6}\frac{d^5u_0}{dx^5} + \cdots \tag{54}$$

By setting $m = 1$ in (53) and (54) and subtracting the resulting equations from (53) and (54), respectively, the following important relations are obtained:

$$\Sigma^{(m)}u_x = \Sigma^{(1)}u_x - \frac{m-1}{2\,m}(u_0 + u_\omega) + \frac{m^2-1}{12\,m^2}\left(\frac{du_0}{dx} - \frac{du_\omega}{dx}\right) - \frac{m^4-1}{720\,m^4}\left(\frac{d^3u_0}{dx^3} - \frac{d^3u_\omega}{dx^3}\right) + \frac{m^6-1}{30240\,m^6}\left(\frac{d^5u_0}{dx^5} - \frac{d^5u_\omega}{dx^5}\right)\cdots \tag{55}$$

$$\Sigma^{(m)}u_x = \Sigma^{(1)}u_x - \frac{m-1}{2\,m}u_0 + \frac{m^2-1}{12\,m^2}\frac{du_0}{dx} - \frac{m^4-1}{720\,m^4}\frac{d^3u_0}{dx^3} + \frac{m^6-1}{30240\,m^6}\frac{d^5u_0}{dx^5} - \cdots \tag{56}$$

These formulas express a relation between the sum of the ordinates separated by unit distance and the ordinates separated by $\frac{1}{m}$th of a unit; the latter sum is expressed in terms of the former, the derivatives at the initial and terminal ordinates, and the number of intervals between unit ordinates made by the interpolated ordinates, namely, m.

Formulas (55) and (56) are commonly referred to as Woolhouse's formulas, probably because Woolhouse[1] published a number of papers in which he developed the formulas and showed their application to important problems in life contingencies.

Examples illustrating the application of the formulas of Lubbock and Woolhouse are given in the *Textbook of the Institute of Actuaries*, Part II, by George King, pp. 467–80, also in *Calculus and Probability*, by Alfred Henry, pp. 114–19. The above example 18 was chosen merely to illustrate the process. The value of these formulas is of course chiefly apparent where the direct calculation of the result would involve excessive labor.

[1] W. S. B. Woolhouse, "On Interpolation, Summation, and the Adjustment of Numerical Tables," *Journal of the Institute of Actuaries*, vol. 11, pp. 61–88, pp. 301–22; vol. 12, pp. 136–76.

"On an Improved Theory of Annuities and Assurances," *Journal of the Institute of Actuaries*, vol. 15, pp. 95–125.

GRADUATION AND SMOOTHING FORMULAS

Graduation by averaging. When observed values are plotted and the ends of the ordinates joined by right lines, the broken polygon is frequently quite irregular. At the same time it may be known from the nature of the observed character that the variation is continuous. In such cases it may be desirable to graduate or smooth the ordinates so that the variations in the broken polygon are not so wide. The population by ages, always overstated at ages which are multiples of 5, is an example of this kind of irregularity.

An obvious method of graduation is to replace each term, u_n, in a series by the average of the term and its two adjacent terms, namely,

$$\tfrac{1}{3}(u_{n-1} + u_n + u_{n+1}).$$

Denoting the process of summing in threes by the operator [3], this average may be written $\frac{1}{3}[3]u_n$. For example,

$$\begin{aligned}[3]u_0 &= u_{-1} + u_0 + u_1.\\ [5]u_6 &= u_4 + u_5 + u_6 + u_7 + u_8.\end{aligned}$$

These operators may be applied in succession in any desired order. Averaging in fives,

$$\begin{aligned}\tfrac{1}{5}[5]u_n &= \tfrac{1}{5}(u_{n-2} + u_{n-1} + u_n + u_{n+1} + u_{n+2})\\ &= \tfrac{1}{5}\{u_n + (u_{n-1} + u_{n+1}) + (u_{n-2} + u_{n+2})\}\\ &= \tfrac{1}{5}(1 + \gamma_1 + \gamma_2)u_n, \qquad (57)\end{aligned}$$

if we define the operator γ_k as follows:

$$\gamma_k u_n = u_{n-k} + u_{n+k}.$$

The graduation may be improved by applying the operator $\frac{1}{5}[5]$ again. The result is

$$\begin{aligned}\tfrac{1}{25}[5]^2u_n &= \tfrac{1}{25}\{5\,u_n + 4(u_{n-1} + u_{n+1}) + 3(u_{n-2} + u_{n+2})\\ &\qquad + 2(u_{n-3} + u_{n+3}) + (u_{n-4} + u_{n+4})\}\\ &= \tfrac{1}{25}(5 + 4\,\gamma_1 + 3\,\gamma_2 + 2\,\gamma_3 + \gamma_4)u_n. \qquad (58)\end{aligned}$$

If the process of averaging is performed a third time, the graduated term becomes

$$\begin{aligned}\tfrac{1}{125}[5]^3u_n = \tfrac{1}{125}\{&19\,u_n + 18(u_{n-1} + u_{n+1}) + 15(u_{n-2} + u_{n+2}) + 10(u_{n-3} + u_{n+3})\\ &+ 6(u_{n-4} + u_{n+4}) + 3(u_{n-5} + u_{n+5}) + (u_{n-6} + u_{n+6})\}\\ = \tfrac{1}{125}(&19 + 18\,\gamma_1 + 15\,\gamma_2 + 10\,\gamma_3 + 6\,\gamma_4 + 3\,\gamma_5 + \gamma_6)u_n. \qquad (59)\end{aligned}$$

It is easy to show that graduation by successive application of the operator $[k]$ will leave an arithmetic progression unchanged.

When first differences are not constant, better results are obtained by employing more powerful smoothing formulas.[1]

Woolhouse, Higham, and Spencer graduation formulas. Woolhouse proposed the first graduation formula of a more general type:

$$G_W = \tfrac{1}{125}(25 + 24\,\gamma_1 + 21\,\gamma_2 + 7\,\gamma_3 + 3\,\gamma_4 - 2\,\gamma_6 - 3\,\gamma_7)u_n. \quad (60)$$

It was derived by assuming u_n to lie on four parabolas of the second degree defined by the sets

$$(u_{n-7},\, u_{n-2},\, u_{n+3}), \qquad (u_{n-6},\, u_{n-1},\, u_{n+4}),$$
$$(u_{n-4},\, u_{n+1},\, u_{n+6}), \qquad (u_{n-3},\, u_{n+2},\, u_{n+7}).$$

The average of u_n and the four values of u_n so obtained was taken as the graduated value. This formula leaves unchanged the terms of a series whose third differences are constant. It employs fifteen terms, seven on each side of the central term, to determine the graduated value of the central term: it can be condensed in its application to the following symbolic form:

$$G_W = \frac{[5]^3}{125}\{10 - 3[3]\{u_n. \quad (61)$$

The graduation formula of Higham involves seventeen terms and is written

$$G_H = \frac{[5]^3}{125}\{[3] - \gamma_2\}u_n. \quad (62)$$

Spencer's twenty-one term formula is perhaps the one most frequently employed when the series to be smoothed contains a large number of terms. Its symbolic form is

$$G_S = \frac{[5]^2[7]}{350}\{1 + [3] - \gamma_3\}u_n. \quad (63)$$

The formulas of Higham and Spencer leave unchanged a series with constant third differences.[2]

The following example of graduation of population statistics of negro females in the original registration states, 1910, by Spencer's formula illustrates the application of these methods.

[1] Corneille L. Landre, *Mathematisch-technische Kapitel zur Lebensversicherung*, second edition, pp. 69–83.

W. S. B. Woolhouse, "Explanation of New Method of Adjusting Mortality Tables," *Journal of the Institute of Actuaries*, vol. 15, pp. 389–410.

[2] J. A. Higham, "On the Adjustment or Graduation of Mortality Tables," *Journal of the Institute of Actuaries*, vol. 23, pp. 335–52.

J. Spencer, "On the Graduation of the Rates of Sickness and Mortality Presented, etc.," *Journal of the Institute of Actuaries*, vol. 38, pp. 334–43.

Table XVI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Age Interval	Population	$\frac{1}{7}(1)$	$[3](2)$	$\gamma_3(2)$	$(2)+(3)-(4)$	$[7](5)$	$\frac{1}{5}(6)$	$[5](7)$	$\frac{1}{10}[5](8)$
15–16	3380	483							
16–17	3862	552	1591						
17–18	3892	556	1786						
18–19	4749	678	1922	1204	1396				
19–20	4819	688	2148	1408	1428				
20–21	5477	782	2191	1459	1514				
21–22	5044	721	2359	1632	1448	11665	2333		
22–23	5991	856	2480	1708	1628	12321	2464		
23–24	6321	903	2713	1675	1941	12802	2560	12562	
24–25	6675	954	2877	1521	2310	12725	2545	12972	
25–26	7140	1020	2867	1835	2052	13299	2660	13278	6512
26–27	6252	893	2713	1697	1909	13716	2743	13269	6505
27–28	5599	800	2672	2035	1437	13851	2770	13042	6398
28–29	6851	979	2573	1530	2022	12754	2551	12490	6186
29–30	5558	794	2854	1603	2045	11592	2318	11898	5908
30–31	7567	1081	2385	1390	2076	10540	2108	11159	5585
31–32	3570	510	2301	1598	1213	10754	2151	10492	5274
32–33	4971	710	1810	1630	890	10155	2031	9810	4991
33–34	4127	590	1919	1652	857	9419	1884	9378	4780
34–35	4332	619	2045	1013	1651	8181	1636	9069	4634
35–36	5856	836	2026	1439	1423	8378	1676	9055	4558
36–37	4001	571	1910	1172	1309	9208	1842	9030	4504
37–38	3521	503	1803	1468	838	10084	2017	9052	4445
38–39	5106	729	1814	1133	1410	9293	1859	8837	4921
39–40	4077	582	2160	1022	1720	8290	1658	8474	4129
40–41	5946	849	1728	844	1733	7305	1461	7817	
41–42	2053	297	1597	1034	860	7393	1479	7105	
42–43	3154	451	1089	1120	420	6799	1360		
43–44	2391	341	1097	1114	324	5736	1147		
44–45	2133	305	1184	563	926				
45–46	3763	538	1108	830	816				
46–47	1855	265	1069	677	657				
47–48	1859	266	910						
48–49	2655	379	981						
49–50	2352	336							

Column (1) is the original population and Column (9) the graduated series. It will be noted that the smoothing process removes the exaggerated population returns at ages 25, 30, 35, and distributes the excess to the other ages. This distribution is such that the quinquennial groups or sums in fives are not greatly changed.

The important investigations in this field are found mostly in actuarial literature, probably because of the many useful applications arising in graduation and smoothing of observed data derived from experience on insured lives. References are given to a number of papers which may fairly be considered as bringing the treatment of this subject up to date.[1]

[1] Robert Henderson, *Graduation of Mortality and Other Tables, Actuarial Studies*, No. 4, pp. 23–42.

George King, "On the Error Introduced into Mortality Tables by Summation Formulas of Graduation," *Journal of the Institute of Actuaries*, vol. 41, pp. 54–96.

George King, "Notes on Summation Formulas of Graduation with Certain New Formulas for Consideration," *Journal of the Institute of Actuaries*, vol. 41, pp. 530–65.

G. J. Lidstone, "On the Rationale of Formulas for Graduation by Summation," *Journal of the Institute of Actuaries*, vol. 41, pp. 348–60, and vol. 42, pp. 106–41.

CHAPTER IV

CURVE-FITTING BY THE METHOD OF LEAST SQUARES AND THE METHOD OF MOMENTS

By E. V. HUNTINGTON

THE PROBLEM OF CURVE-FITTING

Suppose $Y_1, Y_2, \cdots Y_n$ are the ordinates of an empirical curve corresponding to the values $x_1, x_2, \cdots x_n$; and let it be required to find a mathematical curve $y = f(x, a, b, c, \cdots)$ which shall represent the empirical curve as closely as possible.

	Type of Equation	Straight Line	
		Abscissa	Ordinate
(1)	$y = a + bx$	x	Y
(2)	$y = be^{ax}$	x	$\log Y$
(3)	$y = ax^b$	$\log x$	$\log Y$
(4)	$y = a + (b/x)$	$1/x$	Y
(5)	$y = x/(a + bx)$	x	x/Y
(6)	$y = c + be^{ax}$	x	$\log (\Delta Y/\Delta x)$
(7)	$y = c + ax^b$	$\log x$	$\log (\Delta Y/\Delta x)$
(8)	$y = c + b/(x - a)$	$x - x_0$	$(x - x_0)/(Y - Y_0)$
(9)	$y = c + x/(a + bx)$	x	$(x - x_0)/(Y - Y_0)$
(10)	$y = d + cx + be^{ax}$	x	$\log [\Delta^2 Y/(\Delta x)^2]$
(11)	$y = dc^x b^m$, where $m = a^x$	x	$\log [\Delta^2(\log Y)/(\Delta x)^2]$
(12[1])	$y = de^{cx} + be^{ax}$	$(Y_{k+1})/Y_k$	$(Y_{k+2})/Y_k$
(13[1])	$y = e^{ax}(d \cos bx + c \sin bx)$	$(Y_{k+1})/Y_k$	$(Y_{k+2})/Y_k$
(14)	$y = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$	Here the differences of the nth order, $\Delta^n Y$, are constant.	
(15)	The normal curve, Pearson's curves, other frequency curves.	See Chapter VII.	
(16)	Periodic curves: $y = b \sin ax$, etc.	For methods of harmonic analysis, see Running, Lipka, etc.	

[1] In (12) and (13) the values of x are supposed to be equidistant, and Y_k, Y_{k+1}, Y_{k+2} are consecutive values of Y corresponding to x_k, x_{k+1}, x_{k+2}. If $M^2 + 4B$ is positive, use (12); if negative, use (13); where $M =$ slope and $B =$ intercept of the plotted straight line on the Y-axis.

The problem of curve-fitting consists of two parts: (1) selecting the most convenient type of equation $y = f(x, a, b, c, \cdots)$; and (2) determining the parameters, $a, b, c, \cdots$, in the selected equation so that, for a given set of values $x_1, x_2, \cdots x_n$, the computed values, $y_1, y_2, \cdots y_n$, shall agree as closely as possible with the observed values, $Y_1, Y_2, \cdots Y_n$.

The selection of the type of mathematical equation to be used in any given case is usually the most perplexing part of the problem. The preceding table of the most common types may be of assistance.[1]

Plot the points indicated in any row of the table; if these points lie approximately on a straight line, select the equation indicated.

Equations (2) – (14) may also be written in the following straight-line forms, where the quantities in square brackets [ ] suggest the abscissæ and ordinates of the points to be plotted, and the coefficients give the slope of the line and its intercept on the y-axis (see method 2, below).

(2′) $[\log y] = \log b + (a \log e)[x]$.

(3′) $[\log y] = \log a + b[\log x]$.

(4′) $[y] = a + b\,[1/x]$.

(5′) $[x/y] = a + b[x]$.

(6′) $[\log (dy/dx)] = \log (ab) + (a \log e)[x]$.

(7′) $[\log (dy/dx)] = \log (ab) + (b - 1)[\log x]$.

(8′) $\left[\dfrac{x - x_0}{y - y_0}\right] = -\dfrac{a - x_0}{c - y_0} + \dfrac{1}{c - y_0}[x - x_0]$.

(9′) $\left[\dfrac{x - x_0}{y - y_0}\right] = (a + bx_0) + \dfrac{b(a + bx_0)}{a}[x]$.

(10′) $[\log (d^2y/dx^2)] = \log (a^2b) + (a \log e)[x]$ and $[y - be^{ax}] = d + c[x]$.

(11′) $[\log \{d^2(\log y)/dx^2\}] = \log \{(\log b)(\log a)^2/ (\log e)^2\} + (\log a)[x]$

and $[\log y - a^x \log b] = \log d + (\log c)[x]$.

(12′) $\left[\dfrac{y_{k+2}}{y_k}\right] = - e^{(a+c)\Delta x} + (e^{a\Delta x} + e^{c\Delta x})\left[\dfrac{y_{k+1}}{y_k}\right]$

and $[ye^{-cx}] = d + b\,[e^{(a-c)x}]$.

(13′) $\left[\dfrac{y_{k+2}}{y_k}\right] = - e^{2a\Delta x} + (2\, e^{a\Delta x} \cos b\Delta x)\left[\dfrac{y_{k+1}}{y_k}\right]$

and $\left[\dfrac{ye^{-ax}}{\cos bx}\right] = d + c[\tan bx]$.

(14′) (For $y = a + bx + cx^2$). $\left[\dfrac{y - y_0}{x - x_0}\right] = (b + 2\,cx_0) + c[x - x_0]$.

[1] For detailed discussion, see T. R. Running, *Empirical Formulas* (1917), or J. Lipka, *Graphical and Mechanical Computation* (1918).

For the determination of the parameters $a, b, c, \cdots$, after the type of equation has been selected, several methods are available, as follows:

(1) **The method of selected points** is perhaps the simplest, but should never be used except when only the crudest results are required.

A curve with two parameters a, b, can obviously be made to pass exactly through any two selected points, by properly choosing a and b; a curve with three parameters can be made to pass exactly through any three selected points; etc. But the resulting equation may not fit the remaining points at all well, and the choice of the selected points is entirely arbitrary.

(2) **The graphical method** (plotting a straight line as indicated in the table above) is very valuable not only as an aid to the selection of a suitable equation, but also as a means of determining the parameters.

For example, in equation (1) $y = a + bx$, of the table, draw a straight line through the plotted points (estimating the best position by means of a fine thread, or a ruled piece of celluloid), and then measure the slope of the line, and its intercept on the y-axis. The slope gives directly the value of b, and the intercept the value of a.

Again, in equation (2), $y = be^{ax}$, we have $\log_{10} y = \log_{10} b + (.4343\, a)x$. Hence if $\log_{10} y$ is plotted as ordinate against x as abscissa, the slope of the line will give $.4343\, a$, and the intercept of the line will give $\log_{10} b$, whence a and b can at once be computed. (Note that $\log_{10} e = .4343$.)

Similarly in the other cases in the table. This graphical method will give very fair results with a minimum amount of labor.

(3) **The method of averages** determines the slope and intercept of the straight line in question, not by a graphical measurement, but by a simple arithmetical computation.

For example, suppose two parameters, a, b, are to be determined. Substitute the given coördinates in the equation of the straight line, thus obtaining n equations of the first degree in a and b as unknowns. Separate these equations into two approximately equal groups; in each group, add the equations together and divide by their number, thus obtaining two "average" equations of the first degree, from which the two unknowns, a and b, can at once be found. If there are three unknowns, separate the given equations into three groups; etc.

This method of averages is the one most often used in practice. A defect of the method lies in the fact that the manner of grouping the equations is quite arbitrary, and the results will differ somewhat for different groupings.

(4) **The method of least squares** (see below) is the standard method for determining the parameters whenever accurate results are required.

If the selected equation is linear in the parameters, as in case of equations (1), (4), and (14), the method, though laborious, is perfectly straightforward and always yields a definite result.

If the selected equation is not linear in the parameters, but an expansion by Taylor's theorem is possible, then the method will still yield a definite result, by a process of successive approximation (see below); but often only at the cost of excessive labor.

In such cases it is customary to replace the given equation by the equation of the straight line indicated in the table above, which will be linear in certain functions of the parameters; these functions of the parameters are then determined by the method of least squares, and the parameters themselves are computed from these results. It should be noted, however, that this indirect process will not always give the same result as a direct application of the method.

(5) **The method of moments** (see below) is a second systematic method of somewhat wider applicability than the method of least squares, since in many cases where the method of least squares would require extremely laborious successive approximations, the parameters of the selected equation can be expressed explicitly in terms of the moments of the curve.

Even when the process of finding the parameters from the moments requires the solution of a numerical equation by trial and error (as is the case with most of the equations in the table above) the labor is usually much less than if the method of least squares were employed; for numerous examples, see Karl Pearson, " On the Systematic Fitting of Curves," *Biometrika*, vol. 1 (1902), pp. 265–303.

The method is especially useful in the fitting of frequency curves, as will be seen in Chapter VII.

DETERMINATION OF THE PARAMETERS BY THE METHOD OF LEAST SQUARES

According to the " method of least squares," the parameters a, b, c, in the selected equation $y = f(x, a, b, c)$, should be so determined that the *sum of the squares of the " residuals," namely*

$$[f(x_1, a, b, c) - Y_1]^2 + [f(x_2, a, b, c) - Y_2]^2 + \cdots + [f(x_n, a, b, c) - Y_n]^2,$$

(where Y_1, Y_2, $\cdots$ Y_n are the observed values) *shall be a minimum.* It is easily shown that the following working rules will meet this requirement. There are two cases to be considered.

Case 1. Suppose the selected function $f(x, a, b, c)$ is **linear** *in the parameters* a, b, c; that is, suppose

$$f(x, a, b, c) = K + Aa + Bb + Cc,$$

where K, A, B, C are any given functions of x not involving a, b, c.

First, write the n " observation equations "

$$\begin{array}{ll} A_1a + B_1b + C_1c + K_1 - Y_1 = 0, & S_1 \\ A_2a + B_2b + C_2c + K_2 - Y_2 = 0, & S_2 \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot & \cdot \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot & \cdot \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot & \cdot \\ A_na + B_nb + C_nc + K_n - Y_n = 0, & S_n \end{array}$$

where A_1, B_1, C_1, K_1 are the values of the known functions A, B, C, K when $x = x_1$, and Y_1 is the observed value of Y when $x = x_1$; etc.; while a, b, c are to be found. The numbers $S_1, S_2, \cdots S_n$ on the right are check numbers inserted for use later; each S being the value of the expression on the left when $a = 1$, $b = 1$, $c = 1$; thus $S_1 = A_1 + B_1 + C_1 + K_1 - Y_1$.

Next, form the three " normal equations " as follows: (1) Multiply each of the n equations by the coefficient of a in that equation and add; this gives the first normal equation. (2) Multiply each of the n equations by the coefficient of b in that equation and add; this gives the second normal equation. (3) Multiply each of the n equations by the coefficient of c in that equation and add; this gives the third normal equation. In the customary notation:

$$\begin{array}{lll} (1) & [AA]a + [AB]b + [AC]c + [AK] - [AY] = 0. & [AS] \\ (2) & [BA]a + [BB]b + [BC]c + [BK] - [BY] = 0. & [BS] \\ (3) & [CA]a + [CB]b + [CC]c + [CK] - [CY] = 0. & [CS] \end{array}$$

Here the square brackets are used in a special sense, indicating summation, thus:

$$\begin{aligned} [AA] &= A_1^2 + A_2^2 + \cdots + A_n^2; \\ [AB] &= A_1B_1 + A_2B_2 + \cdots + A_nB_n; \\ [AY] &= A_1Y_1 + A_2Y_2 + \cdots + A_nY_n; \text{ etc.} \end{aligned}$$

The check numbers on the right are built up like any of the other coefficients; thus:

$$\begin{aligned} [AS] &= A_1S_1 + A_2S_2 + \cdots + A_nS_n, \\ [BS] &= B_1S_1 + B_2S_2 + \cdots + B_nS_n; \text{ etc.} \end{aligned}$$

If the computation is correct, the check number on the right should equal, in each case, the value of the expression on the left when $a = 1$, $b = 1$, $c = 1$.

Finally, solve the normal equations for a, b, c, according to a systematic schedule which will be sufficiently clear from the example worked out below (for four parameters a, b, c, d). The check numbers on the right, except in the cases of the original equations (1), (2), (3), (4), are built up like any of the other coefficients; the control consists in the fact that in each equation the check number on the right should equal the value of the expression on the left when $a = 1$, $b = 1$, $c = 1$, $d = 1$. By the aid of the check column, the computation may be controlled at the end of every line, or only occasionally, as preferred; or the whole check column may be omitted, except the final control at the bottom of the column.

Example of the solution of normal equations. Given, the four normal equations, (1), (2), (3), (4); to solve for the four unknowns, a, b, c, d.

(1)	$153.000\,a$	$+ 6.285\,b$	$+ 2.485\,c$	$- 27.831\,d$	$- 23.350 = 0$	110.589
(2)	$6.285\,a$	$+ 8.989\,b$	$+ 4.037\,c$	$- 0.426\,d$	$+ 3.697 = 0$	22.582
(3)	$2.485\,a$	$+ 4.037\,b$	$+ 23.616\,c$	$- 3.504\,d$	$+ 9.556 = 0$	36.190
(4)	$- 27.831\,a$	$- 0.426\,b$	$- 3.504\,c$	$+ 9.080\,d$	$- 6.936 = 0$	$- 29.617$
Hence						
(*A* 1)	a	$+ .0410784\,b$	$+ .0162418\,c$	$- .1819020\,d$	$- .1526144 = 0$	$+ .7228039$
(*A* 2)	a	$+ 1.4302307\,b$	$+ .6423230\,c$	$- .0677804\,d$	$+ .5882259 = 0$	$+ 3.5929992$
(*A* 3)	a	$+ 1.6245473\,b$	$+ 9.5034205\,c$	$- 1.4100604\,d$	$+ 3.8454728 = 0$	$+ 14.5633803$
(*A* 4)	a	$+ .0153067\,b$	$+ .1259028\,c$	$- .3262549\,d$	$+ .2492185 = 0$	$+ 1.0641730$
(*A* 1) − (*A* 2)		$- 1.3891523\,b$	$- .6260812\,c$	$- .1141216\,d$	$- .7408403 = 0$	$- 2.8701953$
(*A* 2) − (*A* 3)		$- .1943166\,b$	$- 8.8610975\,c$	$+ 1.3422800\,d$	$- 3.2572469 = 0$	$- 10.9703811$
(*A* 3) − (*A* 4)		$+ 1.6092406\,b$	$+ 9.3775177\,c$	$- 1.0838055\,d$	$+ 3.5962543 = 0$	$+ 13.4992073$
Hence						
(*B* 1)		b	$+ .4506930\,c$	$+ .0821520\,d$	$+ .5333039 = 0$	$+ 2.066149$
(*B* 2)		b	$+ 45.601341\,c$	$- 6.907696\,d$	$+ 16.762577 = 0$	$+ 56.456222$
(*B* 3)		b	$+ 5.8272938\,c$	$- .6734888\,d$	$+ 2.2347524 = 0$	$+ 8.388557$
(*B* 1) − (*B* 2)			$- 45.150648\,c$	$+ 6.989848\,d$	$- 16.229273 = 0$	$- 54.390073$
(*B* 2) − (*B* 3)			$+ 39.774047\,c$	$- 6.234207\,d$	$+ 14.527825 = 0$	$+ 48.067665$
Hence						
(*C* 1)			c	$- .1548117\,d$	$+ .3594472 = 0$	$+ 1.2046355$
(*C* 2)			c	$- .1567406\,d$	$+ .3652589 = 0$	$+ 1.2085183$
(*C* 1) − (*C* 2)				$+ .0019289\,d$	$- .0058117 = 0$	$- .0038828$
Hence						
(*D* 1)				d	$- 3.01296 = 0$	$- 2.01296$

From (*D* 1)	*From* (*C* 1)	*From* (*B* 1)	*From* (*A* 1)	*In* (4) (*check*)
$\therefore d = 3.01296$	$c = + .466441$	$b = - .0482214$	$= + .0340559$	$- 20.4000$
	$- .359447$	$- .2475207$	$- .0017378$	$+ .3532$
	$\therefore c = .106994$	$- .5333039$	$+ .5480634$	$- .3749$
		$\therefore b = - .829046$	$+ .1526144$	$+ 27.3577$
			$\therefore a = .732996$	$- 6.936$
				0.000

Case 2. If the selected function $f(x, a, b, c)$ is **not linear** in a, b, c, we may be able to find the required values of a, b, c, by the following process of *successive approximation;* but the labor of the computation may often become extremely great, or indeed prohibitive. Let a_0, b_0, c_0 be a first approximation, obtained in any manner, and let $\Delta a, \Delta b, \Delta c$ be the "corrections" which must be applied; that is, let

$$a = a_0 + \Delta a, \qquad b = b_0 + \Delta b, \qquad c = c_0 + \Delta c.$$

Then, if we write $K = f(x, a_0, b_0, c_0)$ and

$$A = f'_a\,(x, a_0, b_0, c_0), \qquad B = f'_b\,(x, a_0, b_0, c_0), \qquad C = f'_c\,(x, a_0, b_0, c_0)$$

(where f'_a denotes partial differentiation with respect to a; etc.), we shall have, by Taylor's theorem:

$$f(x, a, b, c) = K + A \cdot \Delta a + B \cdot \Delta b + C \cdot \Delta c + \cdots$$

Since K, A, B, C are functions of x alone, this expression is linear in the corrections $\Delta a, \Delta b, \Delta c$, so that we may build, as in case 1, a set of n "observation equations" for $\Delta a, \Delta b, \Delta c$, leading to the three following "normal equations":

$$\begin{aligned}
[AA]\Delta a + [AB]\Delta b + [AC]\Delta c + [AK] - [AY] &= 0.\\
[BA]\Delta a + [BB]\Delta b + [BC]\Delta c + [BK] - [BY] &= 0.\\
[CA]\Delta a + [CB]\Delta b + [CC]\Delta c + [CK] - [CY] &= 0.
\end{aligned}$$

Solving these equations for $\Delta a, \Delta b, \Delta c$, we have $a = a_0 + \Delta a$, $b = b_0 + \Delta b$, $c = c_0 + \Delta c$, as our second approximations.

Using these values in place of a_0, b_0, c_0, we may repeat the process, thus finding a third approximation; and so on, until the "corrections" become of negligible size.

DETERMINATION OF THE PARAMETERS BY THE METHOD OF MOMENTS

According to the "method of moments," the parameters, a, b, c, should be so determined that as many as possible of the "moments" of the mathematical curve shall equal the corresponding "moments" of the empirical curve; the moments being taken about the y-axis, either as "moments of ordinates" or as "moments of areas." (For illustrative examples, see Chapter VII.)

Moments of ordinates. In rough work, the nth moment of a curve may be taken as the sum of the nth moments of the several ordinates erected at the points $x_1, x_2, \cdots x_n$ in question; the nth moment of an ordinate y_1 being the length of the ordinate multiplied by the nth power of its distance from the y-axis.

Thus, in the case of the mathematical curve $y = f(x, a, b, c)$, the successive moments of the curve, beginning with the zero-th, are

$$m_0 = \Sigma(y_i), \qquad m_1 = \Sigma(x_iy_i), \qquad m_2 = \Sigma(x_i^2y_i), \qquad m_3 = \Sigma(x_i^3y_i), \cdots$$

where the summation extends over $i = 1, 2, 3, \cdots n$.

In the case of the empirical curve, the successive moments of the curve are:

$$M_0 = \Sigma(Y_i), \quad M_1 = \Sigma(x_iY_i), \quad M_2 = \Sigma(x_i^2Y_i), \quad M_3 = \Sigma(x_i^3Y_i), \cdots$$

Hence the equations for determining a, b, c are:

$$\begin{aligned} \Sigma f(x_i, a, b, c) &= \Sigma(Y_i), \\ \Sigma x_if(x_i, a, b, c) &= \Sigma(x_iY_i), \\ \Sigma x_i^2f(x_i, a, b, c) &= \Sigma(x_i^2Y_i). \end{aligned}$$

If the selected function $f(x, a, b, c)$ is linear in a, b, c, these equations can be readily solved. If it is not linear, successive approximations may be used, as in the method of least squares.

(It can be shown that in case the selected curve is a polynomial, $y = a + bx + cx^2 + \cdots$, the method of moments, in the rough form just described, will give precisely the same result as the method of least squares.)

Moments of areas. In refined work it is usual to define the moments, not as moments of ordinates, but rather, as moments of elementary *areas.*

In the case of the mathematical curve, the element of area is obviously ydx, and the successive moments (between the limits $x = x_1$ and $x = x_n$) are given by the definite integrals

$$m_0 = \int ydx, \qquad m_1 = \int xydx, \qquad m_2 = \int x^2ydx, \qquad m_3 = \int x^3ydx \cdots$$

Here m_0 is simply the area under the curve; m_1 is the first, or "static," moment of the area; m_2 is the second moment, or "moment of inertia," of the area; etc.

In the case of the empirical curve, on the other hand, only a finite number of ordinates are given, and it is necessary to agree, first of all, on what shall be taken as the element of area.

The simplest plan is to fill out the curve by constructing a series of rectangles, each of the given ordinates serving as the midordinate of one of the rectangles.[1] On this plan, assuming that the ordinates are

[1] A more refined plan is to fill out the curve, section by section, by the use of some one of the formulas of interpolation (pages 36–45), and then to use the area under each section of this curve in computing the moments M_0, M_1, M_2, M_3, $\cdots$. See Karl Pearson, "On the Systematic Fitting of Curves to Observations and Measurements," *Biometrika*, vol. 1 (1902), p. 265.

equally spaced, at distance dx apart, the successive moments of the empirical curve are

$$M_0=\Sigma(Y_i\Delta x),\ M_1=\Sigma(x_iY_i\Delta x),\ M_2=\Sigma(x_i^2Y_i\Delta x),\ M_3=\Sigma(x_i^3Y_i\Delta x),\cdots$$

The equations for determining a, b, c are then

$$\int f(x,\,a,\,b,\,c)dx = \Sigma(Y_i\Delta x),$$
$$\int xf(x,\,a,\,b,\,c)dx = \Sigma(x_iY_i\Delta x),$$
$$\int x^2f(x,\,a,\,b,\,c)dx = \Sigma(x_i^2Y_i\Delta x).$$

If the selected function $f(x,\,a,\,b,\,c)$ is linear in a, b, c:

$$f(x,\,a,\,b,\,c) = K(x) + aA(x) + bB(x) + cC(x),$$

and if the integrals $\int A(x)dx$, etc., can be found (by mechanical quadrature if necessary), then these equations can be readily solved for a, b, c.

If the function is not linear in a, b, c, recourse may again be had to the method of successive approximation.

Moments about the mean. If the moments of a curve about an arbitrary y-axis are

$$m_0=\int ydx,\qquad m_1=\int xydx,\qquad m_2=\int x^2ydx,\ \cdots,$$

and the moments of the curve about a parallel axis through the center of gravity of the area are

$$\overline{m}_0=\int ydx,\qquad \overline{m}_1=\int(x-\bar{x})ydx,\qquad \overline{m}_2=\int(x-\bar{x})^2ydx,\ \cdots$$

(where $\bar{x}\int ydx=\int xydx$), then the $\overline{m}$'s can be computed from the m's by the following relations:

$\overline{m}_0=m_0.$

$\overline{m}_1=0.$

$\overline{m}_2=m_2-m^2/m_0.$

$\overline{m}_3=m_3-3\,m_1m_2/m_0+2\,m_1^3/m_0^2.$

$\overline{m}_4=m_4-4\,m_1m_3/m_0+6\,m_1^2m_2/m_0^2-3\,m_1^4/m_0^3.$

$\overline{m}_5=m_5-5\,m_1m_4/m_0+10\,m_1^2m_3/m_0^2-10\,m_1^3m_2/m_0^3+4\,m_1^5/m_0^4.$

$$\overline{m}_6=m_6-6\,m_1m_5/m_0+15\,m_1^2m_4/m_0^2-20\,m_1^3m_3/m_0^3+15\,m_1^4m_2/m_0^4-5\,m_1^6/m_0^6.$$

For *Moments of frequency curves*, with *Sheppard's corrections* for *grouping*, etc., see Chapter VII.

CHAPTER V

RANDOM SAMPLING

By H. L. RIETZ

INTRODUCTION

Statistical inferences relating to a class of individuals, say to a large population or race, are very commonly based on observation of part of the population taken at random from the large group. Such a part of the population is called a **random sample.**

Random sampling can perhaps be illustrated most simply by repeated trials in a game of chance. Thus, suppose we make repeated sets of drawings of 100 balls from an urn, always with a constant probability $\frac{1}{4}$ that a ball to be drawn will be white. The number of white balls drawn per set of 100 will fluctuate about 25 as a most probable value (see Bernoulli's theorem, p. 16). These chance fluctuations are often called fluctuations in random sampling. The determination of certain properties of the frequencies obtained in such drawings is one of the simplest problems of the theory of sampling (see p. 72).

We may also illustrate random sampling by considering the determination of the characteristics of a race, say we are to describe height or weight of the adult white male population of a country. It would be unnecessary to measure all the individuals to obtain a high degree of accuracy in the averages. We should almost surely attempt merely to measure an adequate random sample of individuals and to construct our science on the basis of results from the sample. To be sure, the question arises: What is an adequate sample for a particular purpose? The theory of sampling throws some light on this question.

To be more concrete, let us conceive of taking 1000 random samples of 1000 individuals each. Each of these 1000 samples would have its own arithmetic mean, median, standard deviation, and so on. These 1000 arithmetic means would probably differ but slightly from each other in comparison with differences between extreme individuals, but if the measurements are sufficiently accurate, the means would form a frequency distribution. This frequency distribution of means would have its mean (a mean of means) and standard deviation. It is this standard deviation in which we are especially interested, for it gives a measure of the variability of the means of samples.

Along with the above introductory statement about the general meaning of random sampling, it should be emphasized that the existence of a sampling problem in a statistical investigation depends on the purpose of the investigation. On the one hand, there may be no random sampling problem involved because the purpose is merely to describe the facts about the sample. On the other hand, fundamental problems of random sampling arise when the purpose is to predict the properties of an aggregate, or to test a hypothesis about an aggregate by observations on a sample.

THE BERNOULLI THEOREM

Applications of the Bernoulli theorem (p. 16) **to random sampling.** The treatment of random sampling is simplest when we are concerned merely with the fluctuations in frequency where the *a priori* probability p of success in each trial is a constant and where the trials are independent in the probability sense.

Under these conditions, the *a priori* most probable frequency distribution in N sets of n trials each is given by the terms in the expansion of the binomial

$$N(p+q)^n, \text{ where } q = 1 - p. \tag{1}$$

The most probable number, m, of successes in a set of n trials is $m = np$ when np is an integer,[1] and the standard deviation of the number of successes in sets of n trials is $\sqrt{npq}$.

To find the probability that the number of successes n_1 in any set of n trials will fall within a certain conveniently assigned deviation c from the most probable value np, would, in the simplest cases, merely involve evaluating certain terms in the expansion (1), but this method would, in general, be impracticable when n is large. The probability in question is given by the theorem of Bernoulli. In making a random set of n trials with constant probability p of success in each instance, the theorem states that the probability that a deviation $n_1 - np$ in the number of successes will not exceed an assigned number c or that the deviation of the corresponding statistical ratio $\frac{n_1}{n}$ from p, will not exceed $\frac{c}{n}$, is given by

$$P_c = \frac{2}{\sqrt{2\pi npq}}\int_0^c e^{-\frac{x^2}{2npq}}dx + \frac{e^{-\frac{c^2}{2npq}}}{\sqrt{2\pi npq}}. \tag{2}$$

[1] In any case, $np - q \lesseqgtr m \lesseqgtr np + p$.

The application should be restricted to cases where p is not very small in comparison with q (or, of course, q with respect to p). Tables[1] have been prepared which indicate that the formula should hardly be regarded as practicable when $p < .03$.

For the usual purposes of application to statistical problems, the second term in the right-hand member of (2) is so small in comparison with the first term that it may be neglected. Then we have

$$P_c = \frac{2}{\sqrt{2\pi npq}} \int_0^c e^{-\frac{x^2}{2npq}} dx. \qquad (3)$$

This approximation is simply equivalent to the result of assuming a distribution of frequencies in accord with the normal probability function

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}, \qquad (4)$$

where $\sigma = \sqrt{npq}$.

Example 1. In the third edition of *American Men of Science*, by Cattell and Brimhall (1921), p. 804, it is stated that the group of scientific men reported 716 sons and 668 daughters. The data very naturally suggest the old question as to whether these data are at variance with a hypothetical sex ratio $\frac{1}{2}$. In the language of probability, the data suggest the question of simple statistical sampling. In throwing 1384 coins, what is the probability that the number of heads will differ from $\frac{1384}{2} = 692$ by more than $692 - 668 = 24$ on either side?

The probability in question is $1 - P_c$, where we take $c = 24.5$ in (3). To find P_c we may use the table of normal probability integral (table, p. 211), where we find, corresponding to a deviation

$$\frac{x}{\sigma} = \frac{24.5}{\sqrt{npq}} = 1.3171,$$

the value $P_c = .812$. Hence the probability that a random deviation will be greater than 24 is $1 - .812 = .188$. This fact may be expressed roughly by saying that a deviation larger than 24 will occur in the long run slightly more than once per six trials.

Probable error in the number of successes. For a normal distribution of frequencies, the probable error or quartile deviation is that deviation c on either side of the most probable value such that $P_c = \frac{1}{2}$. By reference to the table on page 210, it is found that

$$\frac{c}{\sigma} = \frac{c}{\sqrt{npq}} = .6745,$$

when $P_c = \frac{1}{2}$.

[1] Lucy Whitaker, "On the Poisson Law of Small Numbers," *Biometrika*, vol. 10 (1914), p. 41.

That is, the probable error of the frequency n_1 is

$$PE = .6745\sqrt{npq}, \tag{5}$$

and the probable error of the relative frequency $\frac{n_1}{n}$ is

$$PE = .6745\sqrt{\frac{pq}{n}}. \tag{6}$$

Example 2. Suppose we have 1000 students in a college, of which 250 are women and 750 are men, and that they have been assigned numbers 1, 2, 3, $\cdots$ 1000. Let balls bearing the numbers be placed in an urn from which drawings are to be made at random. Let us draw 100 balls, one at a time at random, under the conditions that each ball is to be replaced so as to keep the probability of drawing a woman's number equal to $\frac{1}{4}$. By (5) the probable error of the number of women's numbers drawn is

$$PE = .6745\sqrt{100(\tfrac{1}{4})(\tfrac{3}{4})} = 2.92.$$

By (6), the probable error of the relative frequency of drawing a woman's number is

$$PE = .6745\sqrt{\frac{(\frac{1}{4})(\frac{3}{4})}{100}} = .0292.$$

We have thus far assumed a constant known probability p. In statistical practice, we are usually obliged to obtain an approximation to p from the data. We then assume that we can get a fair approximation to p by finding a relative frequency of success $p' = \frac{n_1}{n}$, where n is a large number, and n_1 is the number of successes.

Example 3. Suppose we do not know the relative number of men and women in a large group, but we conceive of taking a random sample of 1000 consisting of 300 women and 700 men. We have $\frac{300}{1000}$ as an approximation to p. That is, $p' = .3$. This approximation is subject to the probable error

$$.6745\sqrt{\frac{p'q'}{n}} = .6745\sqrt{\frac{(.3)(.7)}{1000}} = .0098,$$

where

$$q' = 1 - p'.$$

That is, it is an even wager that the true value p for the whole population is between $0.3 - .0098$ and $0.3 + .0098$.

Probability of deviation greater than certain small multiples of *PE*. The chances that the difference between a true value and the value obtained from a sample will exceed numerically, and that it will not exceed numerically an assigned small multiple of the probable error give a convenient form in which to express the probability of certain statistical conclusions. These chances, as obtained from tables of probability functions (pp. 210–16), are approximately as follows:

Assigned Deviation	Probability of Exceeding	Probability of Not Exceeding
PE	.5	.5
2 PE	.177	.823
3 PE	.0430	.9570
4 PE	.00698	.99302
5 PE	.000746	.999254
6 PE	.000052	.999948

These values can easily be expressed in terms of odds in favor of or against assigned deviations (see Chapter VII, p. 100).

In the application of probable error theory or other sampling theory, the fact should be much emphasized that the samples must be chosen in an unbiased manner, otherwise the use of the formulas $\sqrt{npq}$ and $\sqrt{\frac{pq}{n}}$ is invalid.

THE BORTKEWITSCH "LAW OF SMALL NUMBERS"

Application of Poisson's exponential limit — the Bortkewitsch "law[1] of small numbers." As implied on page 73 there is good reason for restricting the applications of the Bernoulli theorem to cases where the probability is not very small, say not smaller than .03. Thus, in considering events which happen rarely (p or q small) we use tables[2] of the Poisson exponential limit in place of the table of the normal probability function in the description of fluctuations in sampling.

Example 4. According to the United States mortality statistics of 1920, there were 9 deaths from measles per 100,000 of population. Let it be required to find the per cent of cases in which the number under random sampling would deviate more than 3 in excess of this number and the per cent of cases in which it would deviate more than 3 in defect of this number.

From *Tables for Statisticians and Biometricians*, p. 122, there would be 12.42 per cent of the cases in excess more than 3, and 11.57 per cent of the cases in defect more than 3. Hence, in 23.99 per cent of the cases, we should get values deviating more than 3 from 9.

From the Bernoulli theorem, we find that 24.33 per cent of the cases would deviate from the most probable by more than 3. Thus, the

[1] L. Von Bortkewitsch, *Das Gesetz der kleinen Zahlen* (1898).

[2] H. E. Soper, "Tables of the Poisson's Exponential Limit," *Biometrika*, vol. 10 (1914), pp. 25–35. See also *Tables for Statisticians and Biometricians* (1914), pp. 113–24.

actual results in percentage of deviation above and below together differ by only .34 per cent, but Poisson's exponential limit shows also the difference between the per cent of cases in excess more than 3 and the per cent in defect more than 3.

THE PROBABLE ERROR

Meaning of probable error in an average or other statistical constant.[1] The conception of fluctuations in sampling applied to simple frequencies and relative frequencies is easily extended to statistical results such as ordinary averages, moment coefficients, correlation coefficients, and so on. Consider the sampling fluctuations of an arithmetic mean as an illustration. For instance, let

$$m_1, m_2, \cdots m_t$$

be the arithmetic means of heights of random samples of t sets of 1000 individuals each of a well-defined class of men. The means

$$m_1, m_2, \cdots m_t$$

will form a frequency distribution whose standard deviation can be found in a form adapted to numerical calculation. This standard deviation could be used as a measure of the fluctuations in sampling of the means. However, instead of using the standard deviation to measure the failure of the mean stature of the sample to agree with the mean stature of the large class of men, it is the usual practice to use the standard deviation multiplied by the constant .6745, and to call this function the probable error of the mean. When the means are normally distributed, this definition of probable error is equivalent to that given on page 73.

We have discussed sampling fluctuations in arithmetic means, but the conception would apply if $m_1, m_2, \cdots m_t$ were any other type of average or statistical constant obtained by random sampling.

Although we have given elsewhere in this Handbook formulas for the probable error in certain statistical averages and coefficients, we shall for convenience of reference now collect together the formulas for the probable error in some of the most important statistical constants.

In the following, n denotes the number of observations; n_1, the number of successes; σ the standard deviation of the observations:

[1] Karl Pearson, "On the Probable Errors of Frequency Constants," *Biometrika*, vol. 2, pp. 273–81.

Statistical Constant	Probable Error
Relative frequency $p' = \frac{n_1}{n}$, $q' = 1 - p'$	$0.6745\sqrt{\frac{p'q'}{n}}$.
Arithmetic mean	$\frac{0.6745\,\sigma}{\sqrt{n}}$.
Median (normal distribution)	$\frac{0.8454\,\sigma}{\sqrt{n}}$.
Standard deviation (normal distribution)	$\frac{0.6745\,\sigma}{\sqrt{2\,n}} = \frac{0.4769\,\sigma}{\sqrt{n}}$.
Quartile (normal distribution)	$\frac{0.9191\,\sigma}{\sqrt{n}}$.
Semi-interquartile range (normal distribution)	$\frac{0.5306\,\sigma}{\sqrt{n}}$.
Coefficient of variation C (normal distribution)	$\frac{0.6745\,C}{\sqrt{2\,n}}\left[1 + 2\left(\frac{C}{100}\right)^2\right]^{\frac{1}{2}} =$ $\frac{0.4769\,C}{\sqrt{n}}\left[1 + 2\left(\frac{C}{100}\right)^2\right]^{\frac{1}{2}}$.
Second moment μ_2 about the mean (normal distribution)	$0.6745\,\sigma^2\sqrt{\frac{2}{n}}$.
Third moment μ_3 about the mean (normal distribution)	$0.6745\,\sigma^3\sqrt{\frac{6}{n}}$.
Fourth moment μ_4 about the mean (normal distribution)	$0.6745\,\sigma^4\sqrt{\frac{96}{n}}$.
Coefficient of correlation r	$0.6745\,\frac{1 - r^2}{\sqrt{n}}$.
Correlation ratio η	$0.6745\,\frac{1 - \eta^2}{\sqrt{n}}$, nearly.
Regression coefficient $r\frac{\sigma_y}{\sigma_x}$	$0.6745\sqrt{\frac{1 - r^2}{n}}\,\frac{\sigma_y}{\sigma_x}$.
Y as computed from the regression equation $y = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$	$0.6745\,\sigma_y\sqrt{1 - r^2}$.
Tetrachoric correlation coefficient	See *Tables for Statisticians and Biometricians*, pp. xl–xlii.
Bi-serial correlation coefficient	See *Biometrika*, vol. 10, pp. 384–390.
$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$, $\beta_2 = \frac{\mu_4}{\mu_2^2}$	See *Tables for Statisticians and Biometricians*, p. lxii.
β_2 for a normal distribution	$0.6745\sqrt{\frac{24}{n}}$.
$\sqrt{\beta_1}$ for a normal distribution	$0.6745\sqrt{\frac{6}{n}}$.

Small samples. In finding probable errors and other measures of fluctuations in sampling, we have assumed that the number of individuals n in a sample is large. These methods are trustworthy when the sample is large, but it is not clear where the boundary is to be drawn between large and small samples. A big problem of practical statistics arises concerning the variation and distribution of averages based on small samples. Considerable progress has been made recently with this problem in relation to the arithmetic mean, standard deviation, and correlation coefficient obtained from small samples. All we shall attempt here is to direct attention to the work of Student,[1] Fisher,[2] and Pearson [3] on this problem.

THE TEST OF GOODNESS OF FIT

Pearson's test of goodness of fit.[4] Given a frequency distribution which represents a random sample taken from a large group. Let us assume that we know the *a priori* most probable frequencies or that we have fitted a theoretical frequency curve to such a distribution by methods of curve-fitting. The question arises as to the goodness of the fit of theory and observation. In considering this question, there is needed a criterion to test whether the theoretical distribution fits the observed distribution well or not.

Pearson's criterion or test of goodness of fit gives a method of determining whether an observed distribution is described by a theoretical curve or distribution to within fluctuations which may reasonably be ascribed to random sampling.

Let us assume that we have n observed frequencies

$$m_1', m_2', \cdots m_n'$$

and n corresponding theoretical frequencies

$$m_1, m_2, \cdots m_n.$$

Pearson's criterion of goodness of fit gives the probability that a series as likely or less likely than the observed series will arise in taking a

[1] Student, "The Probable Error of the Mean," *Biometrika*, vol. 6 (1908), pp. 1–25; "Probable Error of the Correlation Coefficient," *Biometrika*, vol. 6 (1908), pp. 302–10.

[2] R. A. Fisher, "Frequency Distribution of the Values of Correlation Coefficient in Samples from an Indefinitely Large Population," *Biometrika*, vol. 10 (1915), pp. 507–21.

[3] Karl Pearson, "On the Distribution of Standard Deviations of Small Samples," *Biometrika*, vol. 10 (1915), pp. 522–29.

[4] Karl Pearson, "On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling," *Phil. Mag.*, vol. 50, series 5 (1900), pp. 157–75.

random sample. To give a concrete illustration, consider the following simple case of the frequency distribution obtained[1] by throwing 1536 sets of 7 coins per set and noting the number of heads in each throw:

Number of heads . .	0	1	2	3	4	5	6	7
Observed frequency	12	78	270	456	386	252	69	13
Theoretical frequency	12	84	252	420	420	252	84	12

The important problem of sampling is to determine the probability in taking another sample, of obtaining frequencies which deviate as much or more than the observed sample from the theoretical frequencies. To give a precise formulation of the problem, we may state that the extent of deviations is evaluated in terms of a function χ^2 defined below as the sum of the ratios of squares of deviations to corresponding theoretical frequencies.

The method of application of Pearson's criterion of goodness of fit may be stated as follows:

Let
$$\chi^2 = \sum_{t=1}^{t=n}\left\{\frac{(m'_t - m_t)^2}{m_t}\right\}, \qquad (1)$$

where m_t and m'_t are defined above.

The probability that a random sample would exhibit as large or larger deviations from the theoretical than correspond to an assigned χ is

$$P = \sqrt{\frac{2}{\pi}}\int_{\chi}^{\infty} e^{-\frac{\chi^2}{2}} d_{\chi} + \sqrt{\frac{2}{\pi}}\, e^{-\frac{\chi^2}{2}}\left(\frac{\chi}{1} + \frac{\chi^3}{1\cdot 3} + \cdots + \frac{\chi^{n-3}}{1\cdot 3\cdot 5\cdot (n-3)}\right), \qquad (2)$$

if n is even, and

$$P = e^{-\frac{\chi^2}{2}}\left(1 + \frac{\chi^2}{2} + \frac{\chi^4}{2\cdot 4} + \cdots + \frac{\chi^{n-3}}{2\cdot 4\cdot 6\cdot (n-3)}\right), \text{ if } n \text{ is odd.} \qquad (3)^2$$

Example 5. Let us test the agreement of theory and observation as to the number of heads in the results of tossing sets of seven coins, where we have the observed and theoretical frequencies given above in throwing 1536 sets of 7 coins.

$m' =$	12	78	270	456	386	252	69	13
$m =$	12	84	252	420	420	252	84	12
$m' - m =$	0	− 6	18	36	− 34	0	− 15	1
$(m' - m)^2 =$	0	36	324	1296	1156	0	225	1
$\frac{(m' - m)^2}{m} =$	0	.429	1.286	3.086	2.752	0	2.679	.083

Then $\chi^2 = 10.315$.

Turning to Elderton's tables (*loc. cit.*, p. 26) for $n' = 8$, we find $P = .172$ and we

[1] The coins were thrown by students in a class in statistics taught by the writer.

[2] Tables of the values of P by W. P. Elderton are given in *Biometrika*, vol. 1, pp. 155–63, and in *Tables for Biometricians and Statisticians*, pp. 26–29.

conclude that in the long run, approximately 17 times per 100 trials will give deviations as great or greater than those observed.

The theory which underlies the testing of the agreement of theory and observation in statistical practice is usually more complicated than that involved in example 5 because the theoretical values are not *a priori* most probable values, but are determined by fitting a theoretical distribution to the sample. This theoretical distribution is usually some mathematical function suggested by the theory of probability.

Let us assume that we have determined χ^2 and P for such a case by treating the theoretical values obtained by fitting the sample just as if these values were the *a priori* most probable values. In this connection it is fair to say that judgment as to goodness of fit is based on the general order of magnitude of P, and not on small differences in its value. After making use of a certain amount of approximate mathematical analysis Pearson [1] concludes when we are dealing with sufficiently large numbers to give small probable errors that: (1) if a curve is a good fit to a sample as judged by the χ^2 test, to the same fineness of grouping, it may be used to describe other samples of the same population; (2) if the curve is a bad fit to a sample, then the curve cannot serve to the same fineness of grouping to describe other samples from the same population. That is, we have in the χ^2 test a criterion to determine whether a given form of frequency curve describes a sample with a certain degree of fineness of grouping. Furthermore, if the description is good for a certain fineness of grouping, it is good for all rougher groupings. In statistical practice we attempt to get good fitting frequency functions and curves for such groupings as occur in important investigations. Again, it does not seem to hold that a curve fitting a small sample well, will necessarily be a good fit when the number in the sample is greatly increased.

The question naturally arises as to the value of P at which we cease to call a fit good. It is impossible to fix such a value because the changes are gradual. If $P = .01$, the odds are nearly 99 to 1 against a random sample giving as great or greater deviations. If $P = .1$, we anticipate, in the long run, the assigned amount of deviation or more, one time in ten under random sampling. The fit should then surely not be called bad. But it seems undesirable to assign a value of P for which a result must be discarded.

Example 6. The observed frequency distribution of the number of α-particles radiated from a disk in an experiment by Rutherford and Geiger [2] and the correspond-

[1] *Loc. cit.*, pp. 164–66.

[2] *Loc. cit.*, p. 21.

ing theoretical values given by the Poisson exponential limit may be exhibited as follows:

Number of α-particles radiated from a disk in one eighth minute . .	0	1	2	3	4	5	6	7	8	9	10	11	12–14
Observed frequency m'	57	203	383	525	532	408	273	139	45	27	10	4	2
Theoretical frequency m	54	210	407	525	508	394	254	140	68	29	11	4	1
$m' - m$	3	− 7	− 24	0	24	14	19	− 1	− 23	− 2	− 1	0	1
$(m' - m)^2$	9	49	576	0	576	196	361	1	529	4	1	0	1
$\frac{(m' - m)^2}{m}$	.167	.233	1.415	0	1.134	.497	1.421	.007	7.779	.138	.091	0	1

$$\chi^2 = 13.882.$$

Turning to Elderton's tables (*loc. cit.*, p. 27), for $n' = 13$, we find $P = 0.309$. We conclude that, in the long run, approximately 31 cases per 100 trials in taking random samples will give deviations as great or greater than those observed. The inference would surely be that the fit of theory and observation is good in this case.

As a precaution in applying the criterion of goodness of fit in numerical cases, it may be stated that small frequencies at the ends or margins of the distributions should be grouped together as illustrated in example 6 in grouping together the frequencies at 12, 13, and 14. Moreover, the number of class frequencies n should not be very large; for in this case, the test becomes illusory. Fortunately, in practice n is not usually large. As another precaution of a general nature, it should be recognized that the theory of the method is based on the assumption that the deviations in frequencies are a normally distributed system of correlated variables.

The application of the criterion of goodness of fit has been extended to regression curves,[1] to cells in contingency tables [2] and to other problems of sampling.

[1] E. Slutsky, "On the Criterion of Goodness of Fit of Regression Lines and on the Best Method of Fitting Them to Data," *Jour. of Roy. Stat. Soc.*, vol. 77 (1913), pp. 78–84.

[2] R. A. Fisher, "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P," *Jour. Roy. Stat. Soc.*, vol. 85 (1922), pp. 87–94.

G. Udny Yule, "On the Application of the χ^2 Method to Association and Contingency Tables with Experimental Illustrations," *Jour. Roy. Stat. Soc.*, vol. 85 (1922), pp. 95–104.

Karl Pearson, "On the χ^2 Test of Goodness of Fit," *Biometrika*, vol. 14 (1922), pp. 186–91. This paper is a rejoinder to Fisher and Yule.

"Further Note on the χ^2 test of Goodness of Fit," *Biometrika*, vol. 14 (1923), p. 418.

CHAPTER VI

BERNOULLI, POISSON, AND LEXIS DISTRIBUTIONS

By H. L. RIETZ

TYPES OF DISPERSION OF STATISTICAL RATIOS

Relative frequencies or statistical ratios. One of the simplest and most common problems of statistics consists in finding the ratio of the number of cases in which an event happens or in which a condition is fulfilled, to the total number of cases in question. Thus we find the ratio of the number of heads to the number of coins thrown; the ratio of the number of male children born to the total number born; the ratio of the number of deaths or accidents to the population exposed; the ratio of the number of persons under 21 to the total population, and so on. Such statistical ratios or relative frequencies may be obtained for repeated trials in a game of chance, for populations of different districts, or for the same districts at different times, and for a great variety of concrete situations. Experience with actual data shows that such ratios or relative frequencies obtained from different sets of trials exhibit dispersion. It is the purpose of this chapter to present methods for comparing the dispersion of statistical ratios with certain theoretical norms. The method of comparison involves the classification of a statistical series into sub-series for examination as to stability.

The dispersion of relative frequencies or statistical ratios. Certain criteria have been devised[1] for the comparison of dispersion of relative frequencies found from statistical data with the *a priori* most probable value[2] of the dispersion of certain norms derived from urn schemata.

For example, we may compare the dispersion of sex ratios in various districts with the most probable dispersion of the ratio of the number

[1] W. Lexis, *Zur Theorie der Massenerscheinungen in der menschlichen Gesellschaft* (1877), p. 22. "Über die Theorie der Stabilität statistischer Reihen," *Jahrbuch für Nat. Ok. u. Statist.*, vol. 32 (1879), pp. 60–98. *Abhandlungen zur Theorie der Bevölkerungs- und Moralstatistik*, Kap. V–IX, 1903.

[2] The expression "most probable value" means a value more probable than any other single value that can be named.

of heads to the number of coins thrown where the number of coins in a set is equal to the number of births in a district. We may likewise compare the dispersion of death rates of various districts with the dispersion of the binomial distribution $(p + q)^s$, where q is the mean value of the death rates and $p = 1 - q$.

The distributions used as norms are sometimes called:

(1) Bernoulli or binomial distributions.
(2) Poisson distributions.
(3) Lexis distributions.

Bernoulli or binomial distribution.[1] When the probability of an event remains constant throughout any number of sets of drawings, each set consisting of s trials, the resulting distribution is called a **Bernoulli or binomial distribution.**

Thus an urn containing white and black balls is maintained so that in drawing a ball the probability of getting a white ball is p and that of getting a black ball is $q = 1 - p$. *Then the most probable value*[2] *of the arithmetic mean of the number of white balls among s balls taken at random is sp and the most probable value of the standard deviation of the relative frequency of white balls,* $\frac{m}{s}$, *is*

$$\sigma'_B = \sqrt{\frac{pq}{s}}. \tag{1}$$

That is, in drawing s balls one at a time, we get m_1 white and $s - m_1$ black balls, m_2 white and $s - m_2$ black, and so on. Then the *a priori* most probable value of the standard deviation of

$$\frac{m_1}{s}, \frac{m_2}{s}, \ldots$$

is given by (1).

From (1), the most probable value of the standard deviation of the frequencies $m_1, m_2, \cdots$ would clearly be

$$\sigma_B = \sqrt{spq}. \tag{2}$$

Example 1. Conceive of drawing 7 balls one at a time from an urn with a constant probability $\frac{1}{2}$ that a ball will be white. Continue the process by drawing a large number of such sets of 7 balls. Then from (2)

$$\sigma_B = \tfrac{1}{2}\sqrt{7} = 1.323, \tag{3}$$

and the standard deviation of the relative frequency of white balls is

$$\sigma'_B = \tfrac{1}{14}\sqrt{7} = 0.189. \tag{4}$$

[1] G. Udny Yule, *Introduction to the Theory of Statistics*, Chap. 15.

[2] Cf. The Bernoulli Theorem, p. 16.

Experiment 1.[1] A set of 7 coins were thrown 1536 times with the following distribution of heads per throw:

Number of heads .	0	1	2	3	4	5	6	7
Frequencies . . .	12	78	270	456	386	252	69	13

From these frequencies the standard deviation of the number of heads per throw is found to be $\sigma = 1.302$, and the standard deviation of the relative frequency is

$$\sigma' = 0.186. \qquad (5)$$

From (4) and (5), we note the closeness of agreement of theory and experiment. The point to be emphasized in this experiment is that the frequency distribution has at its foundation a constant probability.

Poisson distribution.[2] When the probability of an event varies from trial to trial within a set of s trials, but the several probabilities from one set of s trials are identical with those of every other set, we have a Poisson distribution.

Thus let s urns

$$U_1,\ U_2,\ U_3, \cdots\ U_s$$

contain white and black balls in such relative numbers that

$$p_1,\ p_2,\ p_3, \cdots\ p_s$$

are the probabilities for the respective urns that a ball to be drawn will be white. Let

$$p = \frac{p_1 + p_2 \cdots + p_s}{s}. \qquad (6)$$

Then sp is the most probable value of the arithmetic mean of the number of white balls in a set of s in taking one from each urn, and the standard deviation σ_P of the number of white balls per set of s is related to the standard deviation

$$\sigma_B = \sqrt{spq}$$

of a hypothetical Bernoulli distribution, based on a constant probability p, by the equation

$$\sigma_P{}^2 = spq - \sum_{x=1}^{x=s} (p_x - p)^2. \qquad (7)$$

Hence the standard deviation of a Poisson distribution is less than that of a Bernoulli distribution based on a probability p.

[1] For other experiments which illustrate different types of dispersion, see Arne Fisher, *Theory of Probabilities* (1922), pp. 137–45; also J. M. Keynes, *A Treatise on Probability* (1921), pp. 361–66.

[2] E. Czuber, *Wahrscheinlichkeiterechung*, II (third edition, 1921), pp. 38–40; also Arne Fisher, *loc. cit.* (1922), p. 121; C. V. L. Charlier, *Vorlesungen über die Grundzüge der mathematischen Statistik* (1920), p. 30.

Example 2. Conceive of drawing one ball from each of 7 urns which are maintained so that $\frac{1}{4}$, $\frac{1}{3}$, $\frac{5}{12}$, $\frac{1}{2}$, $\frac{7}{12}$, $\frac{2}{3}$, $\frac{3}{4}$ are the respective probabilities that the balls will be white. Make a record of the number of white balls in the set of 7. Then repeat the process until we know the number of white balls in each of a large number of such sets of 7. Then from (6), we have

$$p = \tfrac{1}{2}.$$

From (7)

$$\sigma_P^2 = \tfrac{14}{9} = 1.556. \tag{8}$$

$$\sigma_P = 1.247. \tag{9}$$

Then the most probable value of the standard deviation σ'_P of the relative frequency of white balls is

$$\sigma'_P = \frac{\sigma_P}{7} = 0.178. \tag{10}$$

Experiment 2. Urn schemata were devised in which each urn contained 12 balls, and in which the number of white balls in the respective urns was 3, 4, 5, 6, 7, 8, 9. A set of 7 balls was obtained by drawing one ball from each urn. The number of white balls in the set of 7 was recorded. Then each ball was returned to the urn from which it was drawn, and a second set of 7 was drawn. This process was continued until we had 480 sets of 7. The following frequency distribution of white balls was obtained from these 480 sets:

Number of white balls .	0	1	2	3	4	5	6	7
Frequency	3	17	75	123	158	83	19	2

From these figures the standard deviation of the number of balls per set is found to be $\sigma = 1.216$, and the standard deviation of the relative frequency of heads is

$$\sigma' = \frac{1.216}{7} = 0.174 \pm 0.004. \tag{11}$$

Note from (10) and (11) the closeness of agreement of theory and experiment. The point to be emphasized is that the frequency distribution of this experiment has at its foundation unequal probabilities within each set of 7, but one set of 7 has at its foundation the same probabilities as any other set of 7.

Lexis distribution. A Lexis distribution is obtained when the probability of an event is constant from trial to trial within a set, but varies from set to set. Thus, we may draw s balls one at a time from an urn U_1 with constant probability p_1 of getting a white ball, from U_2 with a constant probability $p_2 \cdots$, from U_n with constant probability p_n.

Let
$$p = \frac{p_1 + p_2 + \cdots + p_n}{n},$$

where we have n sets of s instances each.

When extended to the n sets of s balls each, the most probable value of the arithmetic mean of the number of white balls, in sets of s balls taken at random, is sp, and the most probable value of the standard deviation of the n numbers giving the white balls in the sets of s is related to the standard deviation $\sqrt{spq}$ of the hypothetical Bernoulli distribution, based on the probability p, by the equation

$$\sigma_L{}^2 = spq + \frac{s^2 - s}{n} \sum_{x=1}^{x=n} (p_x - p)^2 = spq + (s^2 - s)\sigma_{p_x}{}^2, \qquad (12)$$

where σ_{p_x} is the standard deviation of $p_1, p_2, \cdots p_n$.

Hence the standard deviation of a Lexis distribution is greater than that of a Bernoulli distribution based on a constant probability p. In fact, the $\sigma_L{}^2$ may be regarded as the sum of two parts. The first part spq would be present even if the underlying probability were a constant p. Lexis calls this the ordinary or unessential component. The other part $(s^2 - s)\sigma_{p_x}{}^2$, he terms the physical component.

Example 3. Consider 9 urns which are so maintained that $\frac{1}{6}$, $\frac{1}{4}$, $\frac{1}{3}$, $\frac{5}{12}$, $\frac{1}{2}$, $\frac{7}{12}$, $\frac{8}{12}$, $\frac{3}{4}$, $\frac{5}{6}$ are the respective probabilities that a ball taken at random from an urn will be white. Draw 7 balls at random from each urn.

Then the arithmetic mean of the probabilities is $p = \frac{1}{2}$, and the most probable value of the standard deviation of the number of white balls in a set of 7 is given by σ_L in (12). Thus

$$\sigma_L = \frac{\sqrt{133}}{6} = 1.922,$$

and the most probable value of the standard deviation of the relative frequency of white balls in a set of 7 is

$$\sigma'_L = \frac{1.922}{7} = 0.275. \qquad (13)$$

Experiment 3. Urn schemata were arranged in which each of 9 urns contained 12 balls, and the number of white balls in the respective urns was 2, 3, 4, 5, 6, 7, 8, 9, 10. Then 7 balls were drawn one at a time with replacements from each of these urns and a record was made of the number of white balls in each set of 7. This process was repeated 67 times to give large numbers and thus to reduce probable error in our calculation of the standard deviation in the number of white balls in sets of 7. The resulting distribution of white balls in a set of 7 was, for the 603 sets of 7, as follows:

Number of white balls	7	6	5	4	3	2	1	0
Frequencies . . .	32	78	101	97	102	89	73	31

From these figures, we find the standard deviation

$$\sigma = 1.906 \pm .037,$$

and the standard deviation of the relative frequency of balls is

$$\sigma' = \frac{1.906}{7} = 0.272. \qquad (14)$$

From (13) and (14) we note there is a difference of .003 between the experimental result and the *a priori* most probable value. This difference may reasonably be attributed to chance fluctuations. The point to be emphasized in this experiment is that the frequency distribution has at its foundation a constant probability within each set of 7 but unequal probabilities from one set of 7 to another.

This experiment may also be regarded as consisting of 9 sets of 469 instances each, where the drawing of each ball is regarded as an instance. In this case

$$\sigma_L^2 = \frac{469}{4} + \frac{(469)(468)}{9} \cdot \frac{5}{12} = 10,279.$$

$$\sigma_L = 101.4.$$

From the experiment we have the following numbers of balls in 9 sets of 469 balls each:

79, 125, 170, 187, 234, 298, 298, 349, 399.

The standard deviation of these numbers is = 100.4, which agrees well with $\sigma_L = 101.4$ when we allow for chance fluctuations.

THE CRITERIA OF LEXIS AND CHARLIER

The Lexis ratio. Let σ' be the standard deviation of a series of relative frequencies obtained by experiment from statistical data when the probability p is known. Next, find the dispersion

$$\sigma'_B = \sqrt{\frac{pq}{s}},$$

on the hypothesis that we have a Bernoulli distribution.

The ratio $L = \frac{\sigma'}{\sigma'_B} = \frac{\sigma}{\sigma_B}$ is called the **Lexis ratio**, where

$$\sigma = s\sigma' \quad \text{and} \quad \sigma_B = s\sigma'_B.$$

When L[1] $= 1$, the distribution is said to have **normal** dispersion.

When $L < 1$, the distribution is said to have **subnormal** dispersion.

When $L > 1$, the distribution is said to have **supernormal** dispersion.

Thus, in experiment 1, we find from (4) and (5),

$$L = \frac{0.186}{0.189} = 0.984,$$

and the dispersion seems to be approximately normal.

In experiment 2, we find from (4) and (11),

$$L = \frac{0.174}{0.189} = 0.921,$$

and the dispersion seems to be slightly subnormal.

In experiment 3, we find from (4) and (14),

$$L = \frac{0.272}{0.189} = 1.44,$$

and the dispersion is supernormal.

[1] This equality means $L = 1$ except for chance fluctuations.

The probable error of L in these experiments is given by $\pm \dfrac{.4769\,L}{\sqrt{s}}$, but this does not hold when p is estimated from the sample.

The Charlier coefficient of disturbancy. For a Lexis distribution, we have from (2) and (12) the relation

$$\sigma_L^2 = \sigma_B^2 + (s^2 - s)\sigma_{p_x}^2, \tag{15}$$

where σ_{p_x} is the standard deviation of the probabilities from set to set.

From (15),
$$\sigma_{p_x}^2 = \frac{\sigma_L^2 - \sigma_B^2}{s^2 - s}. \tag{16}$$

Based on σ_{p_x} as a measure of variability of probabilities, we may introduce a coefficient of variability

$$\frac{\sigma_{p_x}}{p},$$

where p is the arithmetic mean of the probabilities.

From (16), we have

$$\begin{aligned}\frac{\sigma_{p_x}^2}{p^2} &= \frac{\sigma_L^2 - \sigma_B^2}{(s^2 - s)p^2}\\ &= \frac{\sigma_L^2 - \sigma_B^2}{s^2p^2}\end{aligned} \tag{17}$$

approximately, since we may for the present purpose neglect s in comparison with s^2.

Then
$$\frac{\sigma_{p_x}}{p} = \frac{\sqrt{\sigma_L^2 - \sigma_B^2}}{M}, \tag{18}$$

where $M = sp$ is the most probable value of the mean number of happenings in s trials.

If σ_L in (18) is replaced by the actual standard deviation σ of any given statistical distribution, we have the Charlier **coefficient of disturbancy:**

$$100\,\rho = 100\frac{\sqrt{\sigma^2 - \sigma_B^2}}{M}. \tag{19}$$

The Charlier coefficient $100\,\rho$ is obviously zero for a Bernoulli distribution except for fluctuations in sampling. It is positive for supernormal distributions, and imaginary for subnormal distributions.

APPLICATIONS TO STATISTICAL DATA

Application of the Lexis and Charlier criteria to infantile mortality rates of various districts. As an application of the above criteria, let us consider a set of infantile mortality rates for various States. In the construction of Poisson and Lexis distributions from urn schemata, it is assumed that the number of instances in each set is a constant s. Experiments with urn schemata can be controlled so that this condition is fulfilled, but it is not likely to be fulfilled exactly when we are collecting actual data of the number of births and deaths. The best we can do is to subdivide the total into districts of as nearly equal numbers of births as is practicable, or to weight the rates from different districts in finding means and standard deviations.

Let us consider the dispersion of death rates of white infants under one year of age in the registration of States of the United States.[1]

Before subjecting the data as a whole to analysis, let us consider the death rates of white infants under one year of age in those States in which the number of births of white children is between 33,000 and 58,000. This selection is made so that the number of instances per set has only a moderate amount of variability.

State	Births	Deaths per 1000
California	50,707	70
Connecticut	33,370	85
Indiana	57,915	78
Kansas	35,392	68
Kentucky	53,658	77
Minnesota	51,452	66
North Carolina	51,832	74
Virginia	41,656	78
Wisconsin	54,472	79
Total	430,454	9)675
$AM =$	47,828	75

First Method. The simple AM of the death rates is 75 per thousand, and their standard deviation (without weighting) is 5.72 per thousand.

If these infantile death rates constituted a Bernoulli distribution with a number of instances equal to the average population, 47,828 in each case, we should have

$$\sigma'_B = \sqrt{\frac{pq}{\bar{s}}} = \sqrt{\frac{(.075)(.925)}{47,828}} = .00120 \text{ per person}$$
$$= 1.20 \text{ per thousand.}$$

Hence the Lexis ratio is

$$L = \frac{5.72}{1.20} = 4.77.$$

The Charlier coefficient of disturbancy is

$$100\,\rho = \frac{100\sqrt{(5.72)^2 - (1.20)^2}}{75} = 7.45.$$

Hence the dispersion is supernormal, and the inference is that there is a significant variation in infantile mortality among these States. The interpretation of the result would no doubt involve a consideration of the factors which produce the variation.

[1] *Birth Statistics for the Registration Area of the United States* (1919), p. 37.

Second Method. By weighting the rates with the number of children exposed, we may find the mean death rate per child or the ratio of the total number of deaths (as computed from the death rates) to the total infantile population exposed. Then the rate per thousand children is, for the above illustration,

$$1000\,q = \frac{(50{,}707)(70) + (33{,}370)(85) + (57{,}915)(78) + \cdots + (54{,}472)(79)}{430{,}454}$$
$$= 74.864, \text{ or } .07486 \text{ per child.}$$

Similarly, we may find the standard deviation σ, where

$$\sigma^2 = \frac{50707(70-75)^2 + 33370(85-75)^2 + \cdots + 54472(79-75)^2}{430{,}454} - (75 - 74.864)^2.$$

This gives $\sigma = 5.40$ per thousand.

Based on this method $\sigma'_B = \sqrt{\dfrac{pq}{\bar{s}}} = \sqrt{\dfrac{(.07486)(.92514)}{47{,}828}} = 1.203.$

Hence, by this method, the Lexis ratio is $L = \dfrac{5.40}{1.20} = 4.50.$

It will be noted that the results given by the first method differ somewhat from those given by the second method; but for the usual purpose of drawing inferences about the operation of certain disturbing influences, the two methods would almost surely lead to the same conclusions.[1]

Let us next extend our illustration to include all the States included in the *Birth Statistics for the Registration Area of the United States* (1919), p. 37.

States	Births	Deaths under One Year of Age per 1000 Births
California	50,707	70
Connecticut	33,370	85
Indiana	57,915	78
Kansas	35,392	68
Kentucky	53,658	77
Maine	15,470	91
Maryland	27,448	92
Massachusetts	86,656	87
Michigan	82,876	89
Minnesota	51,452	66
New Hampshire	8,762	93
New York	221,630	82
North Carolina	51,832	74
Ohio	109,652	88
Oregon	13,215	61
Pennsylvania	201,669	98
South Carolina	22,016	76
Utah	12,800	70
Vermont	7,029	86
Virginia	41,656	78
Washington	23,785	62
Wisconsin	54,472	79

[1] For somewhat different methods and for further illustrations, see *Mathematical Theory of Probabilities*, by Arne Fisher (1922), pp. 149–56.

First Method. Arithmetic mean number of births = 57,430.
The simple mean death rate per thousand = 1000 q = 79.545.
The simple standard deviation = 10.24 per thousand, and

$$\sigma'_B = \sqrt{\frac{pq}{s}} = .001129 \text{ or } 1.129 \text{ per thousand.}$$

The Lexis ratio is

$$L = \frac{10.24}{1.129} = 9.07.$$

Second Method. Mean death rate per thousand = 83.049.
Standard deviation = 9.664 per thousand.

$$\sigma'_B = \sqrt{\frac{pq}{s}} = \sqrt{\frac{(.083049)(.916951)}{57430}} = .001151 \text{ per person or } 1.151 \text{ per thousand.}$$

The Lexis ratio is

$$L = \frac{9.664}{1.151} = 8.40.$$

It should be noted that the resulting Lexis ratios differ somewhat by the two methods. However, here again for the usual purpose of passing judgment on the extent of disturbing influences producing departures from a constant probability of death, the conclusions drawn from the two methods would probably not differ significantly.

The next step in the analysis of the given death rates would very naturally be to consider various subsets of States with a view to finding the largest group which may reasonably be regarded as having at its foundation a constant probability of death.

CHAPTER VII

FREQUENCY CURVES

By H. C. CARVER

MODIFICATION OF FREQUENCY MOMENTS

The method usually employed in fitting an arbitrarily chosen frequency function to an observed distribution is the method of moments (see Chapter IV). The process depends on obtaining expressions for as many "*moments*" of the frequency function as there are parameters in the function and assuming that these functional moments may be equated to the numerical moments as computed from the observed distribution. A solution of the resulting equations, theoretically possible, gives the parameters.

In graduating distributions we deal with

(1) observed distributions of either a continuous or discrete variable, and

(2) certain functions of either a continuous or discrete variable.

Five cases may arise.

Case I. In many graduations, the observed distributions are of continuous variables and the functions employed are functions of a continuous variable. Frequencies are represented by areas. (See page 22.)

The functional moments obtained by direct integration are

$$\frac{\int_a x^n y dx}{\int_a y dx}$$

and should be equated to numerical moments of like kind computed from the observed data. But such moments could be computed from the data of the primary series only, since the process of classification in the case of a distribution of a continuous variable gives no clue as to the manner in which the items belonging to the various classes are distributed between their respective class limits.

If we choose the class interval as the unit of x and let x_t be the mid-abscissa of the tth class, then the numerical moments as approximately and directly computed from the observed distribution are in reality defined by

$$N\nu'_n = \sum_{t=-\infty}^{\infty} x_t^n \left\{ \int_{-\frac{1}{2}}^{\frac{1}{2}} f(x_t + h)dh \right\},$$

where N is the total frequency, while the functional moments are

$$N\mu'_n = \int_{-\infty}^{\infty} x^n f(x)dx.$$

The relations between μ'_n and ν'_n, known as Sheppard's adjustments, may be developed as follows:

If $f(x)$ is a continuous function which can be expanded by Taylor's theorem,

$$f(x_t + h) = \sum_{i=0}^{\infty} \frac{h^i}{\underline{|i}} f^{(i)}(x_t),$$

then

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} f(x_t + h)dh = \sum_{i=0}^{\infty} \frac{1}{2^{2i}\underline{|2\,i+1}} f^{(2i)}(x_t),$$

and therefore

$$N\nu'_n = \sum_{t=-\infty}^{\infty} x^n{}_t \left\{ \sum_{i=0}^{\infty} \frac{1}{2^{2i}\underline{|2\,i+1}} f^{(2i)}(x_t) \right\}$$

$$= \sum_{i=0}^{\infty} \frac{1}{2^{2i}\underline{|2\,i+1}} \left\{ \sum_{t=-\infty}^{\infty} x_t^n \cdot f^{(2i)}(x_t) \right\}.$$

If the distribution is such that the theoretical frequency function and all its derivatives vanish at ± ∞, then by the Maclaurin Sum formula,

$$N\nu'_n = \sum_{i=0}^{\infty} \frac{1}{2^{2i}\underline{|2\,i+1}} \left\{ \int_{-\infty}^{\infty} x^n \cdot f^{(2i)}(x)dx \right\}.$$

By successive integration by parts

$$\int_{-\infty}^{\infty} x^n f^{(2i)}(x)dx = n(n-1)\cdots(n - \overline{2\,i-1}) \int_{-\infty}^{\infty} x^{n-2i} f(x)dx$$

$$= N\underline{|2\,i} \cdot {}_nC_{2i} \cdot \mu'_{n-2i}.$$

Therefore

$$\nu'_n = \sum_{i=0}^{\infty} \frac{{}_nC_{2i}}{2^{2i}(2\,i+1)}\,\mu'_{n-2i}$$

$$= \mu'_n + \frac{{}_nC_2}{3 \cdot 2^2}\mu'_{n-2} + \frac{{}_nC_4}{5 \cdot 2^4}\mu'_{n-4} + \cdots \qquad (1)$$

Taking the mean as origin, adjusted moments to as high an order as necessary should be computed from the first set of relations as given below, and the last moment of highest order checked by the proper equation of the second set. Such procedure automatically checks all adjusted moments of inferior order.

From (1), by dropping primes to denote moments about the mean,

$$\begin{aligned}
\mu_2 &= \nu_2 - \tfrac{1}{12}\\
\mu_3 &= \nu_3\\
\mu_4 &= \nu_4 - \tfrac{1}{2}\mu_2 - \tfrac{1}{80}\\
\mu_5 &= \nu_5 - \tfrac{5}{6}\mu_3\\
\mu_6 &= \nu_6 - \tfrac{5}{4}\mu_4 - \tfrac{3}{16}\mu_2 - \tfrac{1}{448}\\
\mu_7 &= \nu_7 - \tfrac{7}{4}\mu_5 - \tfrac{7}{16}\mu_3\\
\mu_8 &= \nu_8 - \tfrac{7}{3}\mu_6 - \tfrac{7}{8}\mu_4 - \tfrac{1}{16}\mu_2 - \tfrac{1}{2304}
\end{aligned} \qquad (2, a)$$

$$\begin{aligned}
\mu_2 &= \nu_2 - \tfrac{1}{12}\\
\mu_3 &= \nu_3\\
\mu_4 &= \nu_4 - \tfrac{1}{2}\nu_2 + \tfrac{7}{240}\\
\mu_5 &= \nu_5 - \tfrac{5}{6}\nu_3\\
\mu_6 &= \nu_6 - \tfrac{5}{4}\nu_4 + \tfrac{7}{16}\nu_2 - \tfrac{31}{1344}\\
\mu_7 &= \nu_7 - \tfrac{7}{4}\nu_5 + \tfrac{49}{48}\nu_3\\
\mu_8 &= \nu_8 - \tfrac{7}{3}\nu_6 + \tfrac{49}{24}\nu_4 - \tfrac{31}{48}\nu_2 + \tfrac{127}{3840}.
\end{aligned} \qquad (2, b)$$

and so on.

It should be noted that Sheppard's corrections are derived on the assumption that the derivatives of the unknown frequency function vanish when $x = \pm\infty$. This does not mean that the derivatives of an arbitrarily chosen graduating function must necessarily vanish at its intermediate range limits. The modification of the observed moments is dependent only on assumptions concerning the vanishing of derivatives of the unknown law of distribution, and these moments are equated to functional moments obtained by integration of a function which contains the graduating function as a factor. The vanishing of derivatives of the graduating function is a detail in finding functional moments, and as such is outside the domain of this chapter.

Case II. In graduating a distribution of a continuous variable by a function of a discrete variable (for example, the point binomial, hypergeometric series, etc.), frequencies should be represented by areas, but ordinates of the function can be computed at certain discrete points only. The frequency areas must be obtained from these ordinates by the use of quadrature formulas.

The computed moments

$$N\nu'_n = \sum_{t=-\infty}^{\infty} x^n{}_t \left\{ \int_{-\frac{1}{2}}^{\frac{1}{2}} f(x_t + h)\,dh \right\}$$

should be adjusted so that they may be equated to the functional moments which are

$$\sum_{t=-\infty}^{\infty} x^n f(x_t).$$

But since, subject to the usual restrictions, the theoretical law of distribution is continuous, and the function vanishes together with all its derivatives at ± ∞, we have

$$\sum_{t=-\infty}^{\infty} x^n f(x_t) = \int_{-\infty}^{\infty} x^n f(x)dx.$$

Therefore the adjustments for Case II are exactly the same as those in Case I.

Case III. If a distribution of a discrete variable be graduated by a function of a continuous variable, and that function, together with all its derivatives, vanishes at the limits of its range, then no adjustments are necessary. This follows from the Maclaurin Sum formula.

Case IV. If a distribution of a discrete variable be graduated by a function of a discrete variable, the numerical and functional moments are at once of the same form, and no adjustments are necessary provided discrete classes are not themselves grouped together to form a new distribution containing fewer classes.

Case V. It should be noted that for a distribution of a continuous variable the graduated frequencies should be obtained by computing areas by either direct integration or quadrature formulæ. It is sometimes expedient to regard the frequencies of distributions of continuous variables as though the frequencies were associated with discrete variables. On this assumption no modification of moments would be called for (see Cases III and IV) and the ordinates of the graduating function may be taken as representing the graduated frequencies.

Such procedure must be regarded as theoretically unsound, since frequencies of a distribution of continuous variables should be represented by areas. Practically, the results obtained will differ but little by the two methods except in the case where class frequencies lie in the abrupt region of extremely skew distributions.

PROBABLE ERRORS OF FREQUENCY MOMENTS

This subject, which is of prime importance to an understanding of the theory of frequency distributions, has been treated in considerable detail in papers[1] by Pearson, Filon, and Sheppard.

[1] W. F. Sheppard, "On the Application of the Theory of Error to Cases of Normal Distribution and Normal Correlation," *Phil. Trans. A.*, vol. 192 (1898), pp. 101–67.

K. Pearson and L. N. G. Filon, "On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation," *Phil. Trans. A.*, vol. 191 (1898), pp. 229–311.

K. Pearson, "On the Probable Errors of Frequency Constants," *Biometrika,* vol. 2 (1903), pp. 273–81.

In our treatment of frequency distributions we shall have occasion to deal with the following formulas relating to probable errors.

(*a*) The standard deviation of errors in the *s*th moment taken about the mean is

$$\sigma_{\mu_s} = \sqrt{\frac{\mu_{2s} - \mu_s{}^2 + s^2\mu_2\mu_{s-1}{}^2 - 2\,s\mu_{s-1}\mu_{s+1}}{n}}.$$

(*b*) If $\alpha_s = \frac{\mu_s}{\sigma^s}$,

$$\sigma_{\alpha_s} =$$

$$\sqrt{\frac{1}{n}\left\{\alpha_{2s} - \alpha_s{}^2 + s^2\alpha^2{}_{s-1} - 2\,s\alpha_{s-1}\alpha_{s+1} + \frac{s^2}{4}\alpha_s{}^2(\alpha_4 - 1) - s\alpha_s(\alpha_{s+2} - \alpha_s - s\alpha_3\alpha_{s-1})\right\}}.$$

(*c*) If $\beta_1 = \frac{\mu_3{}^2}{\mu_2{}^3}$, then $\beta_1 = \alpha_3{}^2$, $\sigma_{\beta_1} = 2\,\alpha_3\sigma_{\alpha_3}$,

and from (*b*),

$$\sigma_{\beta_1} = \alpha_3\sqrt{\frac{1}{n}\{4(\alpha_6 - 6\,\alpha_4 + 9) + \alpha_3(35\,\alpha_3 + 9\,\alpha_3\alpha_4 - 12\,\alpha_5)\}}.$$

(*d*) If $\beta_2 = \frac{\mu_4}{\mu_2{}^2}$, *i.e.* $\beta_2 = \alpha_4$,

then from (*b*)

$$\sigma_{\beta_2} = \sqrt{\frac{1}{n}\{\alpha_8 + \alpha_4(4\,\alpha_4{}^2 - \alpha_4 - 4\,\alpha_6) + 8\,\alpha_3(2\,\alpha_3 + 2\,\alpha_3\alpha_4 - \alpha_5)\}},$$

where the probable error in any statistical result Z may be defined as equal to 0.6745 σ_Z (cf. p. 76).

THE NORMAL CURVE

Referred to the mean of a distribution as origin, the equation of the normal curve is

$$y = \frac{N}{\sigma\sqrt{2\,\pi}}e^{-\frac{x^2}{2\,\sigma^2}},$$

where N and σ represent the total frequency and the standard deviation of the distribution (cf. p. 13). This curve is symmetrical about the centroidal axis $x = 0$, and y is a maximum at $x = 0$; consequently the mean, median, and mode coincide.

This equation has been developed from various hypotheses. Hagen's demonstration[1] is based on the hypothesis that any observed variation from the mean is the algebraic result of an indefinitely large number of elementary variations which are of equal infinitesimal magnitude and each of which is equally likely to be positive or negative.

[1] Mansfield Merriman, *Method of Least Squares* (1910), p. 17.

Czuber's *Wahrscheinlichkeitsrechnung*[1] gives a very careful development which follows essentially the memoirs of M. W. Crofton.[2]

It is desirable to measure deviations from the mean in units of the standard deviation of the distribution in order that the distribution itself may be analyzed quite independently of the unit of measurement involved.

Placing, therefore,

$$t = \frac{x}{\sigma},$$

the equation of the normal curve reduces to

$$y = \frac{N}{\sigma}\phi_{(t)},$$

where

$$\phi_{(t)} = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}.$$

It follows, therefore, that if the ordinates of the frequency polygon associated with any observed distribution be multiplied by $\frac{\sigma}{N}$ and $f(t)$ be defined by $f(t) = \frac{\sigma}{N}y = \frac{\sigma}{N}f(x)$, the plotted points (t, f_t) will fall approximately on the graph of $\phi_{(t)}$ provided the distribution be nearly "normal." The distribution (t, f_t) is referred to as a "standard distribution."

Table I exhibits the reduction of two apparently symmetrical distributions to their standard distributions. The representation of (t, f_t) as coördinates of points in a plane and the corresponding normal curves $\phi(t)$ would indicate at a glance that each of the curves closely follows the normal law, although it would be difficult to see which of the distributions more nearly approximates the law. For more exact criteria we may proceed as follows:

For the normal curve

$$\alpha_{2s} = \frac{\mu_{2s}}{\sigma^{2s}} = \frac{\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} x^{2s}e^{-\frac{x^2}{2\sigma^2}}dx}{\sigma^{2s}}$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} t^{2s}e^{-\frac{t^2}{2}}dt = 1\cdot 3\cdot 5\cdots(2s-1) = \frac{\lfloor 2s}{2^s\lfloor s}.$$

Since $\phi_{(t)}$ is symmetrical, that is, an even function of t,

$$\alpha_{2s+1} = 0.$$

[1] *Ableitung des Fehlergesetzes aus der Hypothese der Elementarfehler*, vol. 1 (1914), p. 291.

[2] *Phil. Trans.* (1868), and "Probability," *Encyclopædia Britannica* (1885), vol. 19.

Table I

	Distribution I			Distribution II		
x	f_x	t	f_t	f_x	t	f_t
− 10	2	− 3.8957	.0006	9	− 3.6803	.0023
− 9	4	− 3.5069	.0012	14	− 3.3149	.0036
− 8	14	− 3.1181	.0042	35	− 2.9494	.0090
− 7	41	− 2.7294	.0123	64	− 2.5840	.0164
− 6	83	− 2.3406	.0249	122	− 2.2186	.0312
− 5	169	− 1.9519	.0506	240	− 1.8531	.0614
− 4	394	− 1.5631	.1181	505	− 1.4877	.1291
− 3	669	− 1.1744	.2005	783	− 1.1222	.2002
− 2	990	− .7856	.2966	1198	− .7568	.3063
− 1	1223	− .3969	.3664	1444	− .3914	.3693
0	1329	− .0081	.3982	1584	− .0259	.4051
1	1230	.3806	.3685	1496	.3395	.3826
2	1063	.7694	.3185	1242	.7050	.3176
3	646	1.1582	.1936	888	1.0704	.2271
4	392	1.5469	.1175	568	1.4358	.1452
5	202	1.9357	.0605	283	1.8013	.0724
6	79	2.3244	.0237	135	2.1667	.0345
7	32	2.7132	.0096	53	2.5322	.0136
8	16	3.1019	.0048	23	2.8976	.0059
9	5	3.4907	.0015	5	3.2630	.0013
10	2	3.8794	.0006	8	3.6285	.0020
11				1	3.9939	.0003
12				1	4.3594	.0003

	Distribution I	Distribution II
Σf_x	8585	10701
$\Sigma x f_x$	179	759
$\Sigma x^2 f_x$	56809	80183
b	.02085032 [1]	.070927951
ν'_2	6.6172394	7.4930380
ν_2	6.6168047	7.4880072
σ	2.5723150	2.7364223
$1/\sigma$	.38875488	.36544067
σ/N	.00029962900	.00025571650

[1] For the meaning of b, see page 24.

From (*b*), page 96, we may therefore obtain the following values of α_s together with their probable errors:

TABLE II. CRITERIA FOR THE NORMAL CURVE OF ERROR

s	α_s	PROBABLE ERROR OF α_s
3	0	$.67449\sqrt{6/n} = 1.652/\sqrt{n}$
4	3	$.67449\sqrt{24/n} = 3.304/\sqrt{n}$
5	0	$.67449\sqrt{720/n} = 18.10/\sqrt{n}$
6	15	$.67449\sqrt{6120/n} = 52.77/\sqrt{n}$
7	0	$.67449\sqrt{124110/n} = 237.6/\sqrt{n}$
8	105	$.67449\sqrt{1663200/n} = 869.9/\sqrt{n}$

For distributions I and II we compute:

	DISTRIBUTION I		DISTRIBUTION II	
s	μ_s	α_s	μ_s	α_s
2	6.5334714	1.0000	7.4046739	1.0000
3	− .20783947	− .012445491	− 1.4969639	− .074293774
4	134.40994	3.1487878	179.68978	3.2772645
5	− 9.5158173	− .087213982	− 107.92005	− .72333137
6	4828.4828	17.313254	7914.1779	19.493418
7	− 369.01209	− .51765046	− 6056.2890	− 5.4819595
8	243628.25	133.70623	400330.38	133.16644

On the hypothesis that these distributions are normal, the difference between the actual and expected values of α_s may be compared with the probable error of α_s as follows:

	DISTRIBUTION I			DISTRIBUTION II		
s	(1) Actual — Expected	(2) P. E. of α_s	(3) = (1) ÷ (2)	(1) Actual — Expected	(2) P. E. of α_s	(3) = (1) ÷ (2)
3	− .012445	.01783	.698	− .072938	.01597	4.57
4	.14879	.03566	4.17	.27726	.03194	8.68
5	− .087214	.1953	.447	− .72333	.1750	4.13
6	2.3133	.5695	4.06	4.4934	.5101	8.81
7	− .51765	2.564	.202	− 5.4820	2.297	2.39
8	28.706	9.389	3.06	28.166	8.409	3.35

Since, according to the normal hypothesis, the odds in favor of a variable lying between

± P. E. =	1 to 1
± 2 P. E. =	4.5 to 1
± 3 P. E. =	21 to 1
± 4 P. E. =	142 to 1
± 5 P. E. =	1310 to 1
± 6 P. E. =	19200 to 1
± 7 P. E. =	420000 to 1
± 8 P. E. =	17000000 to 1
± 9 P. E. =	1000000000 to 1

it is evident that the hypothesis that distribution II is normal is untenable. The first distribution is more nearly normal and certainly exhibits no significant skewness.

TABLE III

x	t Mid-Class Interval	ϕ_t	Graduated Frequencies	Ungraduated Frequencies
− 12	− 4.6732	.00001		
− 11	− 4.2844	.00004		
− 10	− 3.8957	.00020	1	2
− 9	− 3.5069	.00085	3	4
− 8	− 3.1181	.00309	10	14
− 7	− 2.7294	.00963	32	41
− 6	− 2.3406	.02578	86	83
− 5	− 1.9519	.05937	198	169
− 4	− 1.5631	.11759	392	394
− 3	− 1.1744	.20018	668	669
− 2	− .7856	.29302	978	990
− 1	− .3969	.36872	1231	1223
0	− .0081	.39892	1331	1329
1	.3806	.37106	1238	1230
2	.7694	.29673	990	1063
3	1.1582	.20400	681	646
4	1.5469	.12059	403 [1]	392
5	1.9357	.06128	205	202
6	2.3244	.02678	89	79
7	2.7132	.01005	34	32
8	3.1019	.00325	11	16
9	3.4907	.00090	3	5
10	3.8794	.00021	1	2
11	4.2682	.00004		
12	4.6570	.00001		
Total		2.57230	8585	8585

[1] Actual value 402.46, nearest integral value 402. If 402 be used, the total frequency would amount to 8584, instead of 8585, due to the fact that nearest integral values have been used in all instances. In such cases we modify that frequency which is nearest the desired boundary.

The statement that the higher moments, because of their large probable errors, are idle, is frequently made but not always true. The ratio of a variation to its probable error is the deciding factor, and in the above illustration the higher moments reflect information quite as significant as that presented by the moments of lower order.

With the aid of the table, page 209, distribution I, treated as a normal distribution of a discrete variable, may be graduated as in Table III by using

$$\sigma = \sqrt{\nu_2},$$

as explained in Case III of the discussion of the modification of frequency moments.

Regarded as a distribution of a continuous variable (Case I)

$$\sigma = \sqrt{\mu_2} = \sqrt{\nu_2 - \tfrac{1}{12}}$$

and one may proceed as in Table IV.

Table IV

Mid-Class Interval	Lower Limits of Class Interval		$\int_{-\infty}^{t} \phi(t)\,dt$	Unit Frequencies	Frequencies	
x	x	t			Graduated	Un-graduated
− 12	− 12.5	− 4.8985	.00000			
− 11	− 11.5	− 4.5073	.00000	.00002		
− 10	− 10.5	− 4.1160	.00002	.00008	1	2
− 9	− 9.5	− 3.7248	.00010	.00033	3	4
− 8	− 8.5	− 3.3336	.00043	.00120	10	14
− 7	− 7.5	− 2.9424	.00163	.00374	32	41
− 6	− 6.5	− 2.5511	.00537	.01002	86	83
− 5	− 5.5	− 2.1599	.01539	.02308	198	169
− 4	− 4.5	− 1.7687	.03847	.04572	393	394
− 3	− 3.5	− 1.3774	.08419	.07783	668	669
− 2	− 2.5	− .9862	.16202	.11390	978	990
− 1	− 1.5	− .5950	.27592	.14333	1230	1223
0	− .5	− .2038	.41925	.15512	1332	1329
1	.5	.1875	.57437	.14423	1238	1230
2	1.5	.5787	.71860	.11535	990	1063
3	2.5	.9699	.83395	.07931	681	646
4	3.5	1.3611	.91326	.04689	403	392
5	4.5	1.7524	.96015	.02381	204	202
6	5.5	2.1436	.98396	.01042	89	79
7	6.5	2.5348	.99438	.00391	34	32
8	7.5	2.9260	.99829	.00125	11	16
9	8.5	3.3173	.99954	.00036	3	5
10	9.5	3.7085	.99990	.00008	1	2
11	10.5	4.0997	.99998	.00002		
12	11.5	4.4909	1.00000			
Total				1.00000	8585	8585

Table V

Height, without Shoes, Inches	Ungraduated Frequencies	Graduated as Distribution of a	
		Continuous Variable	Discrete Variable
55 –		.02	.02
56 –		.14	.14
57 –	2	.67	.67
58 –	4	2.84	2.84
59 –	14	10.30	10.30
60 –	41	32.11	32.11
61 –	83	86.03	86.03
62 –	169	198.18	198.17
63 –	394	392.44	392.43
64 –	669	668.13	668.11
65 –	990	977.93	977.92
66 –	1223	1230.60	1230.63
67 –	1329	1331.38	1331.41
68 –	1230	1238.38	1238.41
69 –	1063	990.33	990.33
70 –	646	680.88	680.86
71 –	392	402.46	402.44
72 –	202	204.52	204.51
73 –	79	89.35	89.35
74 –	32	33.56	33.56
75 –	16	10.83	10.84
76 –	5	3.01	3.01
77 –	2	.72	.72
78 –		.15	.15
79 –		.03	.03
Total	8585	8584.99	8584.99

In order to illustrate the discussion of Case V, Table V is presented, showing the results which would have been obtained from Tables III and IV had more extensive tables been employed. Actually the distribution is one of a continuous variable, but the results obtained by regarding it as one of a discrete variable, and consequently omitting the modification of moments, differ but very little from the results obtained by properly treating it as a distribution of a continuous variable.

It should be borne in mind that according to the criteria it is doubtful whether or not this distribution is normal. Distributions satisfying the criteria are rare. But if the ratio of the difference between the actual and expected values to the probable error is not greater than five or six, the results of the graduation are in general sufficiently satisfactory to justify the procedure on practical, if not on theoretical, grounds.

After completing the graduation of the distribution by means of the normal curve, the Pearson χ^2-test of goodness of fit (see p. 78), gives

$$P = .268,$$

if we use for the theoretical frequencies the nearest integers to the graduated frequencies shown in Table V. This value of P is somewhat larger than would probably be obtained if frequencies near the end of the range were grouped together (see p. 81).

In order to provide for situations where the use of normal curve is not permissible, an entirely different function or a more general law must be employed.

PEARSON'S GENERALIZED FREQUENCY CURVES

Certain geometrical properties of unimodal frequency distributions suggest that any associated frequency function may be represented as a solution of the differential equation

$$\frac{dy}{dt} = \frac{y(a-t)}{f(t)}, \tag{1}$$

since

(*a*) if there be one mode only, there must be a value of $t = a$ for which the derivative vanishes, and

(*b*) towards the extremes, as y approaches zero, the derivative must also approach zero.

At present we place no restriction on $f(t)$ except that we assume that it may be expanded in a converging power series. Equation (1) may then be written as

$$\frac{1}{y}\frac{dy}{dt} = \frac{a-t}{b_0 + b_1t + b_2t^2 + \cdots}, \tag{2}$$

where the mean of the distribution is taken as origin, and the abscissæ are measured in units of the standard deviation, as in the discussion of the normal curve of error.

Clearing (2) of fractions, multiplying through by t^n, and integrating over the range r to s with respect to t gives

$$\left[a\int t^n y\,dt - b_0\int t^n dy - b_1\int t^{n+1}dy - \cdots - \int t^{n+1}y\,dt\right]_{t=r}^{s} = 0. \tag{3}$$

But

$$\left[\int t^n dy\right]_{t=r}^{s} = \left[t^n y - n\int t^{n-1}y\,dt\right]_{t=r}^{s}$$

and if the frequency function, when multiplied by t^n, vanishes at the limits of the distribution, r and s, we have that

$$\int_{t=r}^{s} t^n dy = -nN\alpha_{n-1}, \tag{4}$$

where α_n is defined on page 96.

Giving n successively the values 0, 1, 2, · · ·, we obtain from (3) and and (4), noting that N, the total frequency, cancels out and that $\alpha_0 = 1$, $\alpha_1 = 0$, $\alpha_2 = 1$,

$$\begin{array}{rcrcrcrcl} a & + & & & b_1 & & & + \cdots = & 0 \\ & & b_0 & + & & & 3\,b_2 & + \cdots = & 1 \\ a & + & & & 3\,b_1 & + & 4\,\alpha_3 b_2 & + \cdots = & \alpha_3 \\ \alpha_3 a & + & 3\,b_0 & + & 4\,\alpha_3 b_1 & + & 5\,\alpha_4 b_2 & + \cdots = & \alpha_4 \\ \alpha_n a & + & n\alpha_{n-1} b_0 & + & (n+1)\alpha_n b_1 & + & (n+2)\alpha_{n+1} b_2 & + \cdots = & \alpha_{n+1} \end{array} \tag{5}$$

If we assume that $f(t)$ converges so rapidly that terms involving the third and higher powers of t may be neglected, a simultaneous solution of (5) yields

$$(6) \quad \begin{cases} a = -\dfrac{\alpha_3}{2(1+2\,\delta)}, & b_1 = \dfrac{\alpha_3}{2(1+2\,\delta)}, \\ b_0 = \dfrac{2+\delta}{2(1+2\,\delta)}, & b_2 = \dfrac{\delta}{2(1+2\,\delta)}, \end{cases}$$

where
$$\delta = \frac{2\,\alpha_4 - 3\,\alpha_3^2 - 6}{\alpha_4 + 3},$$

and the moments are defined by the recurring relation

$$\alpha_{n+1} = \frac{n}{2-(n-2)\delta}\left[(2+\delta)\alpha_{n-1} + \alpha_3\alpha_n\right]. \tag{7}$$

The value of
$$a = -\frac{\alpha_3}{2(1+2\,\delta)},$$
which represents the distance between the mean and mode expressed in standard units, is called by Pearson the " skewness " of the distribution.

If $\alpha_3 = \delta = 0$, then by (6) the differential equation

$$\frac{1}{y}\frac{dy}{dt} = \frac{a-t}{b_0 + b_1 t + b_2 t^2}$$

reduces to

$$\frac{1}{y}\frac{dy}{dt} = -t,$$

which on integration yields the normal curve of error,

$$y = y_0 e^{-\frac{t^2}{2}},$$

which has already been discussed. Pearson refers to this curve as Type VII, a special case of his generalized frequency curves.

If $\delta = 0$ but $\alpha_3 \neq 0$, we obtain the differential equation

$$\frac{1}{y}\frac{dy}{dt} = -\frac{\frac{\alpha_3}{2} + t}{1 + \frac{\alpha_3}{2}t},$$

yielding after integration Pearson's Type III,

$$y = y_0\left(1 + \frac{\alpha_3}{2}t\right)^{\frac{4}{\alpha_3^2} - 1} e^{-\frac{2}{\alpha_3}t}. \tag{8}$$

To determine the constant of integration impose the condition that the total frequency must equal N and it follows that

$$N = y_6\int_{-\frac{\alpha_3}{2}}^{\infty}\left(1 + \frac{\alpha_3}{2}t\right)^{\frac{4}{\alpha_3^2} - 1} e^{-\frac{2}{\alpha_3}t}dt.$$

Placing

$$\frac{2}{\alpha_3}\left(\frac{2}{\alpha_3} + t\right) = z,$$

the above reduces to

$$\begin{aligned} N &= y_0\frac{\alpha_3}{2}e^n\left(\frac{1}{n}\right)^{n-1}\int_0^{\infty} e^z z^{n-1}dz \\ &= y_0\frac{e^n}{n^{n-\frac{1}{2}}}\Gamma_{(n)} \end{aligned}$$

or

$$y_0 = \frac{Nn^{n-\frac{1}{2}}}{e^n\Gamma_{(n)}} \tag{9}$$

where

$$n = \frac{4}{{\alpha_3}^2}\cdot$$

But since

$$\Gamma_{(n)} = \sqrt{2\pi}\,n^{n+\frac{1}{2}}e^{-n+\frac{1}{12n}-\frac{1}{360n^3}\cdots}$$

(9) reduces to

$$y_0 = \frac{N}{\sqrt{2\pi}}\left[1 - \frac{{\alpha_3}^2}{48} + \frac{{\alpha_3}^4}{4608} - \cdots\right]. \tag{9, a}$$

In the above we have taken the standard deviation as the unit of the independent variable. If the distribution be plotted with the class interval as the unit, the value of (9, a) must be divided by σ and we have

$$y_0 = \frac{N}{\sigma\sqrt{2\pi}}\left[1 - \frac{{\alpha_3}^2}{48} + \frac{{\alpha_3}^4}{4608} - \cdots\right] \tag{9, b}$$

which, for purposes of computation, is superior to (9) inasmuch as it does not require the use of a Gamma Function, and is so very rapidly converging that ordinarily one need go no further than the term $\frac{\alpha_3^2}{48}$. For $\alpha_3 = 0$ this constant and Type III curve obviously reduce to the Type VII functions.

Moments for distributions which obey this law are governed by the recurring relation

$$\alpha_{n+1} = n\left[\alpha_{n-1} + \frac{\alpha_3}{2}\alpha_n\right] \qquad (10)$$

which is obtained from (7) by placing $\delta = 0$.

This equation yields the following values of the functional moments, placing $\frac{\alpha_3^2}{2} = \gamma$:

$$\begin{aligned} \alpha_4 &= 3[1 + \gamma], \\ \alpha_5 &= 2\,\alpha_3[5 + 3\,\gamma], \\ \alpha_6 &= 5[3 + 13\,\gamma + 6\,\gamma^2], \qquad (11) \\ \alpha_7 &= 3\,\alpha_3[35 + 77\,\gamma + 30\,\gamma^2], \\ \alpha_8 &= 7[15 + 170\,\gamma + 261\,\gamma^2 + 90\,\gamma^3], \\ &\text{etc.} \end{aligned}$$

The probable error of α_s which is equal to $.67449\,\sigma_{\alpha_s}$ may be obtained from the following relations:

$$\begin{aligned} n\sigma^2_{\alpha_3} &= 6 + 18\,\gamma + \tfrac{15}{2}\,\gamma^2, \\ n\sigma^2_{\alpha_4} &= 24 + 504\,\gamma + 1002\,\gamma^2 + 378\,\gamma^3, \\ n\sigma^2_{\alpha_5} &= 720 + 14400\,\gamma + 54000\,\gamma^2 + 60750\,\gamma^3 + 18630\,\gamma^4, \quad (12) \\ n\sigma^2_{\alpha_6} &= 6120 + 328800\,\gamma + 2289150\,\gamma^2 + 5157600\,\gamma^3 \\ &\qquad + 4363830\,\gamma^4 + 1158300\,\gamma^5, \\ &\text{etc.} \end{aligned}$$

If a distribution may be represented properly by Type III, the value of b_2 in the equation

$$\frac{1}{y}\frac{dy}{dt} = \frac{a - t}{b_0 + b_1 t + b_2 t^2}$$

should be small, — theoretically equal to zero. Actually the value of b_2 as computed from (6) will rarely if ever exactly equal zero, and the question as to whether or not the computed value is significant may be answered by comparing its value with the probable error b_2. This probable error, on the assumption that the distribution does belong to Type III, may be computed from

$$\sigma_{b_2}^2 = \frac{8 + 60\,\gamma + 84\,\gamma^2 + 27\,\gamma^3}{3(2 + \gamma)^2 n}. \tag{13}$$

From the relations (11), (12), and (13), Tables VI and VII may be computed. With the aid of these tables one can readily compute for this curve from the value of α_3, by interpolation, the expected higher moments and their probable errors, and also the probable error of b_2.

TABLE VI. TYPE III. TABLE OF α_s FOR CERTAIN VALUES OF α_3

α_3	α_s				
	$s = 4$	$s = 5$	$s = 6$	$s = 7$	$s = 8$
.0	3.000	.000	15.000	.000	105.000
.1	3.015	1.003	15.326	10.616	110.996
.2	3.060	2.024	16.312	21.931	129.536
.3	3.135	3.081	17.986	34.673	162.307
.4	3.240	4.192	20.392	49.622	212.215
.5	3.375	5.375	23.594	67.641	283.527
.6	3.540	6.648	27.674	89.698	382.069
.7	3.735	8.029	32.726	116.898	515.481
.8	3.960	9.536	38.872	150.509	693.529
.9	4.215	11.187	46.246	191.986	928.475
1.0	4.500	13.000	55.000	243.000	1235.500

TABLE VII. TYPE III. TABLE OF $.67449\,\sigma_z\sqrt{n}$ FOR CERTAIN VALUES OF α_3

α_3	$.67449\,\sigma_z$						
	$z = \alpha_3$	$z = \alpha_4$	$z = \alpha_5$	$z = \alpha_6$	$z = \alpha_7$	$z = \alpha_8$	$z = b_2$
.0	1.652	3.304	18.098	52.77	237.6	869.9	.5507
.1	1.665	3.475	18.998	59.65	274.5	1106.8	.5596
.2	1.701	3.961	21.648	78.81	381.0	1771.3	.5858
.3	1.762	4.709	25.974	108.84	559.4	2915.1	.6279
.4	1.846	5.680	31.990	150.67	828.4	4756.4	.6841
.5	1.951	6.858	39.829	206.97	1221.1	7660.9	.7526
.6	2.077	8.244	49.727	281.64	1784.3	12163.9	.8317
.7	2.223	9.848	62.001	379.56	2580.1	19033.0	.9201
.8	2.387	11.688	77.033	506.72	3689.4	29341.4	1.0167
.9	2.570	13.783	95.259	670.23	5215.8	43634.5	1.1203
1.0	2.771	16.156	117.171	878.51	7289.7	66697.9	1.2302

For the distribution of Table VIII one finds

s	ν_s	μ_s	α_s
2	8.6907427	8.6074094	1.000
3	14.721596	14.721596	.58297
4	278.29904	273.98284	3.69754
5	1488.6512	1476.3832	6.79232
6	19353.480	19009.385	29.8093
7	175579.80	172989.69	92.4627
8	2171577.4	2126981.9	387.502

Also

$$\delta = .05607,$$
$$b_2 = .0252 \pm .0079.$$

Since the value of b_2 lies between three and four times its probable error, it appears that the values of α_3 and α_4 do not support the hypothesis that this distribution belongs to Type III, although the evidence that it does not belong to Type III is not conclusive. By straight-line interpolation one can obtain the approximate comparison of the actual and expected values of α_s ($s = 4, 5, 6, 7, 8$) for (a) $\alpha_3 = .583$ and (b) $\alpha_3 = .6$ as follows:

s	Observed α_s (1)	$\alpha_3 = .583$		$\alpha_3 = .6$		
		Expected α_s (2)	(3) = (1) − (2)	Expected α_s (4)	(5) = (1) − (4)	P. E. α_s (6)
4	3.70	3.51	.19	3.54	.16	.08
5	6.79	6.43	.36	6.65	.14	.48
6	29.8	27.0	2.8	27.7	2.1	2.72
7	92.5	85.9	6.6	89.7	2.5	17.25
8	388.	365.	23.	382.	6.0	118.

For $\alpha_3 = .58297$, which has a probable error of ± .035, it is noted that the expected higher moments are for each value of s smaller than the corresponding observed moments. But from formulas (11) it is noted that γ is always positive and moreover that an increase in any value of α_3 will automatically increase the values of all expected higher moments. In other words, since α_3 is subject to a probable error just as all other moments, one is not justified in exactly reproducing it at the expense of all other moments. Thus if we select $\alpha_3 = .6$ as the α_3 of the law of distribution, we note that a much better agreement between the higher moments is obtained. A slightly higher value of α_3 would give still closer values for the higher moments, but this improvement would be gained at the expense of $\alpha_3 = .583 \pm .035$.

The addition of the b_2t^2 term of the differential equation, giving rise to Pearson's other generalized frequency curves, will not necessarily remedy the situation. Thus, from the recurring relation of equation (7) we may compute the higher moments obtaining the following results:

s	Actual α_s	Expected α_s	Actual — Expected	P. E. α_s
5	6.79	7.11	− .32	.80
6	29.8	32.1	− 2.3	4.7
7	92.5	112.5	− 20.	38.4
8	387.	535.	− 148.	359.

Table VIII

Ungraduated	x	$1 + .3t$	log $(1 + .3t)$	$\frac{91}{9}$ log $(1 + .3t)$	$-\frac{10t}{3}$ log e	log y	Graduated y
3	− 8	.11584	− .93614	− 9.46542	4.26650	− 2.04137	
9	− 7	.21761	− .66232	− 6.69679	3.77544	.23620	2
46	− 6	.31937	− .49571	− 5.01218	3.28438	1.42975	27
167	− 5	.42113	− .37558	− 3.79753	2.79332	2.15334	142
372	− 4	.52290	− .28158	− 2.84709	2.30226	2.61272	410
718	− 3	.62466	− .20436	− 2.06631	1.81120	2.90244	799
1186	− 2	.72642	− .13881	− 1.40352	1.32014	3.07417	1186
1462	− 1	.82819	− .08187	− .82780	.82908	3.15873	1441
1498	0	.92995	− .03154	− .31890	.33802	3.17667	1502
1460	1	1.03171	− .01355	.13701	− .15304	3.14152	1385
1142	2	1.13348	.05442	.55025	− .64410	3.06370	1158
913	3	1.23524	.09174	.92759	− 1.13516	2.94998	891
642	4	1.33701	.12613	1.27531	− 1.62622	2.80664	641
435	5	1.43877	.15800	1.59756	− 2.11728	2.63783	434
235	6	1.54053	.18766	1.89745	− 2.60834	2.44666	280
167	7	1.64230	.21545	2.17844	− 3.09940	2.23659	173
133	8	1.74406	.24157	2.44254	− 3.59046	2.00963	102
47	9	1.84582	.26618	2.69138	− 4.08152	1.76741	59
29	10	1.94759	.28950	2.92717	− 4.57258	1.51214	33
13	11	2.04935	.31163	3.15093	− 5.06364	1.24484	18
9	12	2.15112	.33266	3.36356	− 5.55470	.96641	9
5	13	2.25288	.35274	3.56659	− 6.04576	.67838	5
8	14	2.35464	.37192	3.76052	− 6.53682	.38125	2
2	15	2.45641	.39030	3.94637	− 7.02788	.07604	1
	16	2.55817	.40793	4.12463	− 7.51894	− .23776	1
10701							10701

$\Sigma y = 10701$ $\quad \nu'_1 = .688347$

$\Sigma xy = 7366$ $\quad \nu'_2 = 9.164564$

$\Sigma x^2y = 98070$ $\quad \nu_2 = 8.6907427$ $\quad y = y_0(1 + .3\,t)^{\frac{91}{9}}e^{-\frac{10}{3}t}$

$\sigma = 2.9480065$

$\log y_0 = 3.15755$

The differences between the actual and expected moments is considerably greater for $s = 7$ and 8, although the functional higher moments tend to minimize the seriousness of the variations.

The questions

(*a*) precisely how many terms in the denominator of the differential equation should be retained?

(*b*) are we justified in absolutely reproducing certain moments at the expense of others?

(*c*) what is the best criterion for goodness of fit?

may be regarded at the present time as furnishing material for research.

Table VIII shows the graduation of the distribution on the hypothesis that the distribution belongs to Type III, treating $\alpha_3 = .6$ and further treating it as one of a discrete variable. The moments are therefore unmodified.

If graduated as a distribution of a continuous variable, the moments must be modified and the graduated frequencies — being areas — approximated by a quadrature formula.

If it be impossible to graduate a particular distribution by either the normal curve or Type III, the equation

$$\frac{1}{y}\frac{dy}{dt} = \frac{a - t}{b_0 + b_1t + b_2t^2}$$

must be used.

The integration of this equation depends upon the nature of the roots of the quadratic $b_0 + b_1t + b_2t^2$. Thus if the roots be real, the integration of the equation yields

$$y = y_0\left(1 + \frac{t}{\alpha_1}\right)^{m_1}\left(1 + \frac{t}{\alpha_2}\right)^{m_2}, \tag{14}$$

and if the roots be complex

$$y = y_0\left(1 + \frac{t^2}{a^2}\right)^{m_1} e^{m_2 tan^{-1}\frac{t}{a}}. \tag{15}$$

For an extended discussion of these curves, including special and transitional cases, the following sources may be consulted.

(*a*) "Mathematical Contributions to the Theory of Evolution, II," by Karl Pearson, *Phil. Trans.*, A (1895), vol. 185, pp. 343 *et seq.*

(*b*) *Frequency Curves and Correlation*, by W. P. Elderton. C. and E. Layton, London, 1906.

(*c*) *A First Course in Statistics*, by D. Caradog Jones. G. Bell and Sons, Ltd., London, 1921.

(*d*) "Mathematical Contributions to the Theory of Evolution." Second supplement to a memoir on Skew Variations, by Karl Pearson, *Phil. Trans.*, A (1916), vol. 216, pp. 429–57.

DIFFERENCE EQUATION GRADUATION

The reasoning which prompted Pearson to choose his *differential* equation also suggests

$$\frac{\Delta y_x}{\Delta x} = \frac{y_x(a - x)}{b_0 + b_1x + b_2x^2 + \cdots} \tag{16}$$

as the *difference* equation of a unimodal distribution.

By arbitrarily allowing Δx to represent the difference in magnitudes of two successive class marks, this element may be considered equal to unity.

It may be noted that a determination of the values of the constants of the differential equation (2) does not permit a graduation of a distribution until the equation has been integrated, yielding a solution of the form $y = y_0 \cdot f(x)$. In using the difference equation the situation is otherwise, for as soon as the values of

$$\frac{y_{x+1}}{y_x} = 1 + \frac{\Delta y_x}{y_x}$$

are computed, these ratios permit a calculation of a series of ordinates which are proportional to those required of the graduation. The condition that the sum of the graduated ordinates must equal those of the ungraduated will always determine the proper factor of proportionality.

Therefore, since the difference equation requires no integration in the course of the graduation the idea of type — an outgrowth of integration — is of secondary importance.

Writing (16) as

$$(b_0 + b_1x + b_2x^2 + \cdots)\Delta y_x = (a - x)y_x, \tag{17}$$

multiplying through by x^n and summing with respect to x yields

$$b_0\Sigma x^n\Delta y_x + b_1\Sigma x^{n+1}\Delta y_x + b_2\Sigma x^{n+2}\Delta y_x + \cdots = a\Sigma x^n y_x - \Sigma x^{n+1}y_x. \tag{18}$$

If the range of the distribution be from $x = -\infty$ to $x = \infty$ and we assume that the extreme ordinates y_x and also $x^n y_x$ vanish, then

$$\Sigma x^n\Delta y_x = -{}_nC_1\Sigma x^{n-1}y_x + {}_nC_2\Sigma x^{n-2}y_x - {}_nC_3\Sigma x^{n-3}y_x + \cdots$$

Giving n in (18) successively the values 0, 1, 2, $\cdots$, we obtain corresponding to (5), selecting the mean as origin,

$$\begin{array}{rrrrl}
a & + & b_1 & - & b_2 + \cdots = 0, \\
 & b_0 & - b_1 & + & (3\nu_2+1)b_2 + \cdots = \nu_2, \\
\nu_2 a & - b_0 + & (3\nu_2+1)b_1 + & & (4\nu_3 - 6\nu_2 - 1)b_2 + \cdots = \nu_3, \\
\nu_3 a + & (3\nu_2+1)b_0 + & (4\nu_3 - 6\nu_2 - 1)b_1 + & & (5\nu_4 - 10\nu_3 + 10\nu_2 - 1)b_2 + \cdots = \nu_4
\end{array}$$

where

$$\nu_n = \frac{\Sigma x^n y_x}{\Sigma y_x}. \tag{19}$$

If we now set

$$\alpha_n = \frac{\nu_n}{\sigma^n}$$

and

$$\delta' = \frac{2\,\alpha_4 - 3\,\alpha_3^2 - 6 + \dfrac{1}{\nu_2}}{\alpha_4 + 3 - \dfrac{1}{\nu_2}}, \tag{20}$$

a solution of (19) yields — keeping no terms in $b_0 + b_1x + b_2x^2 + \cdots$ of greater than the second degree —

$$(21)\left\{\begin{aligned} a &= \frac{-\nu_3}{2\,\nu_2(1 + 2\,\delta')} - \frac{1}{2}, & b_1 &= b_2 - a,\\ b_0 &= -a + \frac{\nu_2(2 + \delta')}{2(1 + 2\,\delta')}, & b_2 &= \frac{\delta'}{2(1 + 2\,\delta')}.\end{aligned}\right.$$

Table IX

Ungraduated Frequencies	Abscissa Measured from		$x^2 + c_1x + c_2$	$x^2 + c_3x + c_4$	$\dfrac{y_{x+1}}{y_x}$	Proportional y_x	Graduated y_x
	Arbitrary Origin	Mean $= x$					
3	− 8	− 8.6883	243.87	15.812	15.4238	10	1
9	− 7	− 7.6883	235.73	38.269	6.1597	154	7
46	− 6	− 6.6883	229.58	62.727	3.6600	950	43
167	− 5	− 5.6883	225.44	89.185	2.5277	3477	158
372	− 4	− 4.6883	223.29	117.64	1.8980	8789	401
718	− 3	− 3.6883	223.14	148.10	1.5067	16683	760
1186	− 2	− 2.6883	225.00	180.56	1.2461	25136	1145
1462	− 1	− 1.6883	228.85	215.02	1.0643	31322	1427
1498	0	− .6883	234.71	251.47	.9333	33338	1519
1460	1	.3117	242.56	289.93	.8366	31115	1418
1142	2	1.3117	252.41	330.39	.7640	26031	1186
913	3	2.3117	264.27	372.85	.7088	19887	906
642	4	3.3117	278.12	417.31	.6665	14096	642
435	5	4.3117	293.98	463.76	.6339	9394	428
235	6	5.3117	311.83	512.22	.6088	5955	271
167	7	6.3117	311.68	562.68	.5895	3625	165
133	8	7.3117	353.54	615.14	.5747	2137	97
47	9	8.3117	377.39	669.60	.5636	1228	56
29	10	9.3117	403.25	726.05	.5554	692	32
13	11	10.3117	431.10	784.51	.5495	384	18
9	12	11.3117	460.95	844.97	.5455	211	10
5	13	12.3117	492.81	907.43	.5431	115	5
8	14	13.3117	526.66	971.89	.5419	63	3
2	15	14.3117	562.52	1038.34	.5417	34	2
	16	15.3117	600.37	1106.80	.5424	18	1
	17	16.3117	640.23	1177.26	.5438	10	
	18	17.3117	682.08	1249.72	.5458	5	
10701						234859	10701

It follows that we may therefore write

$$\frac{y_{x+1}}{y_x} = \frac{x^2 + c_1x + c_2}{x^2 + c_3x + c_4}, \tag{22}$$

where

$$\begin{aligned} c_1 &= -1 - \left(1 - \frac{\nu_3}{\nu_2}\right)\frac{1}{\delta'}, & c_3 &= c_1 + 4 + \frac{2}{\delta'}, \\ c_2 &= \nu_2\left(1 + \frac{2}{\delta'}\right), & c_4 &= c_2 + c_3 - 1. \end{aligned} \tag{23}$$

Table IX illustrates difference equation graduation. The results are practically identical with those obtained by using Pearson's generalized curve IV, which is the curve which would be used according to Pearson's criteria.

Graduation of the stump of a distribution. From equation (18) we obtain the equations

$$(24) \quad \begin{cases} b_0\Sigma\ \ \Delta y_x + b_1\Sigma x\,\Delta y_x + b_2\Sigma x^2\Delta y_x = a\Sigma\ \ y_x - \Sigma x\,y_x, \\ b_0\Sigma x\,\Delta y_x + b_1\Sigma x^2\Delta y_x + b_2\Sigma x^3\Delta y_x = a\Sigma x\,y_x - \Sigma x^2y_x, \\ b_0\Sigma x^2\Delta y_x + b_1\Sigma x^3\Delta y_x + b_2\Sigma x^4\Delta y_x = a\Sigma x^2y_x - \Sigma x^3y_x, \\ b_0\Sigma x^3\Delta y_x + b_1\Sigma x^4\Delta y_x + b_2\Sigma x^5\Delta y_x = a\Sigma x^3y_x - \Sigma x^4y_x. \end{cases}$$

To graduate the stump of a distribution, or even an entire distribution, these equations may be used directly without restricting the origin to the mean.

If it be desired to graduate that portion of a distribution lying between $x = r$ and $x = s$, the summations must be from $x = r$ to $x = s - 1$, since Δy_{s-1} involves y_s. Clearly we could not sum to $x = s$, since such procedure would require a knowledge of the frequency at $x = s + 1$, which is contrary to hypothesis.

It is not necessary to compute by the long method two sets of moments; that is, $\Sigma x^n y_x$ and $\Sigma x^n \Delta y_x$ since

$$\begin{aligned} \Sigma\ \ \Delta y_x &= y_s - y_r \\ \Sigma x\,\Delta y_x &= (s-1)\,y_s - (r-1)\,y_r - \Sigma y_x \\ \Sigma x^2\Delta y_x &= (s-1)^2y_s - (r-1)^2y_r - 2\,\Sigma x\,y_x + \Sigma yx \\ \Sigma x^3\Delta y_x &= (s-1)^3y_s - (r-1)^3y_r - 3\,\Sigma x^2y_x + 3\,\Sigma xy_x - \Sigma y_x \\ \Sigma x^4\Delta y_x &= (s-1)^4y_s - (r-1)^4y_r - 4\,\Sigma x^3y_x + 6\,\Sigma x^2y_x - 4\,\Sigma xy_x + \Sigma y_x \\ \Sigma x^5\Delta y_x &= (s-1)^5y_s - (r-1)^5y_r - 5\,\Sigma x^4y_x + 10\,\Sigma x^3y_x - 10\,\Sigma x^2y_x \\ &\qquad + 5\,\Sigma xy_x - \Sigma y_x \end{aligned}$$

Further examples of graduations of frequency distributions and stumps of frequency distributions by the use of difference equations may be found in a paper by the writer.[1]

[1] H. C. Carver, "On the Graduation of Frequency Distributions," *Proceedings of the Casualty Actuarial and Statistical Society of America*, vol. 6, Part I, No. 13, pp. 52–72.

THE GENERALIZED NORMAL CURVE — CHARLIER THEORY

The fact that many distributions do not satisfy "normal" criteria has led to many generalizations of the normal curve.[1] Of such generalizations, that of Charlier is particularly noteworthy. In *Ueber das Fehlergesetz*[2] and *Die Zweite Form des Fehlergesetzes*[2] Charlier has shown that on the hypothesis of elementary errors any law of error may assume one of the following two forms:

Type A.

$$F(x) = A_0\phi_{(x)} + A_3\phi_{(x)}^{(3)} + A_4\phi_{(x)}^{(4)} + \cdots, \tag{1}$$

where

$$\phi_{(x)} = \frac{1}{\sigma\sqrt{2\,\pi}}\,e^{-\frac{(x-b)^2}{2\,\sigma^2}}.$$

Type B.

$$F(x) = B_0\psi_{(x)} + B_1\Delta\psi_{(x)} + B_2\Delta^2\psi_{(x)} + \cdots, \tag{2}$$

where $\psi_{(x)} = \dfrac{e^{-\lambda}\sin\pi x}{\pi}\left[\dfrac{1}{x} - \dfrac{\lambda}{(x-1)\underline{|1}} + \dfrac{\lambda^2}{(x-2)\underline{|2}} - \dfrac{\lambda^3}{(x-3)\underline{|3}} + \cdots\right].$

If x be a positive integer or zero, the above series in x reduces to

$$\psi_{(x)} = \frac{e^{-\lambda}\lambda^x}{\underline{|x}}.$$

In *Ueber die Darstellung willkürlicher Functionen*[3] the values of the coefficients A_n, B_n, which are independent of x are obtained by imposing the usual conditions associated with the method of moments.

The function

$$y = \frac{N}{\sigma\sqrt{2\,\pi}}\left[1 - \frac{\mu_3}{2\,\sigma^3}\left(\frac{x}{\sigma} - \frac{1}{3}\frac{x^3}{\sigma^3}\right)\right]e^{-\frac{x^2}{2\,\sigma^2}},$$

which has been stressed by Bowley and Edgeworth is equivalent to the first two terms of the type A function as noted above.

For purposes of computation the type A curve may be written in the form

$$F(t) = \frac{N}{\sigma}\left[\phi_{(t)} - \frac{c_3}{\underline{|3}}\phi_{(t)}^{(3)} + \frac{c_4}{\underline{|4}}\phi_{(t)}^{(4)} - \frac{c_5}{\underline{|5}}\phi_{(t)}^{(5)} + \cdots\right], \tag{3}$$

where

$$\begin{aligned}
c_3 &= \alpha_3 \\
c_4 &= \alpha_4 - 3 \\
c_5 &= \alpha_5 - 10\,\alpha_3 \\
c_6 &= \alpha_6 - 15\,\alpha_4 + 30 \\
c_7 &= \alpha_7 - 21\,\alpha_5 + 105\,\alpha_3 \\
c_8 &= \alpha_8 - 28\,\alpha_6 + 210\,\alpha_4 - 315
\end{aligned}$$

[1] See Chapter III, Part II, Bowley (P. S. King & Son, London, 1920), and *Law of Error*, by Edgeworth, *Camb. Phil. Trans.* (1904), vol. 20.

[2] *Arkiv för Matematik, Astronomi och Fysik*, Band 2. N: O 8.

[3] *Loc. cit.*

and in general

$$c_n = \alpha_n - \frac{n^{(2)}}{2}\alpha_{n-2} + \frac{n^{(4)}}{2^2\lfloor 2}\alpha_{n-4} - \frac{n^{(6)}}{2^3\lfloor 3}\alpha_{n-6} + \frac{n^{(8)}}{2^4\lfloor 4}\alpha_{n-8}, \text{ etc.}$$

If series (3) is so rapidly convergent that terms after the third or fourth may be neglected, it affords us a simple representation of any distribution.

The rapidity with which this series converges, however, depends upon the extent to which the generating function, $\phi(t)$, is a fair approximation of the unknown law of distribution. In general, if the distribution be approximately symmetrical, equation (3) is very rapidly convergent. If, however, the distribution be bimodal or extremely asymmetrical, series (3) as a function of the symmetrical generating function, $\phi(t)$, is for practical purposes of no value because

(*a*) the number of required terms in (6) would necessitate an impracticable amount of labor; and

(*b*) the probable errors of α_n for values of n greater than 5 or 6 are generally so large that the assumption that we may substitute the moments computed from the observed data for the theoretical moments is open to serious criticism.

It should be understood, however, that theoretically equation (3) is capable of representing any frequency function associated with graduated variates, but that convergence means one thing and rapid convergence quite another.

Table IX *a* shows the results of graduating a distribution by using $\phi(t)$ as a generating function. Table IX *b* shows the successive steps in the computation of the graduated values of the function by the use of three terms of the series (3). Values of the integral $\int_0^t \phi(t)dt$, and of $\phi(t)$ and its derivatives are given in tables on pp. 209–216.

In case series (3) is not sufficiently rapidly convergent, we may proceed as follows:

Let θ_t be any arbitrarily chosen function which may be used as a statistical generating function and which, being a frequency function, may be expressed by means of equation (3) in the series

$$\theta_t = N'\left\{\phi(t) - \frac{d_3}{\lfloor 3}\phi^{(3)}(t) + \frac{d_4}{\lfloor 4}\phi^{(4)}(t) - \frac{d_5}{\lfloor 5}\phi^{(5)}(t) + \cdots\right\}, \qquad (4)$$

where d_n is a function of the moments of θ_t.

Eliminating $\phi(t)$ between (3) and (4) we obtain

$$F(t) = \frac{N}{N'}\left\{\theta_t - \frac{A'_3}{\underline{|3}}\theta_t^{(3)} + \frac{A'_4}{\underline{|4}}\theta_t^{(4)} - \frac{A'_5}{\underline{|5}}\theta_t^{(5)} + \cdots\right\}, \tag{5}$$

where

$$\begin{aligned} A'_3 &= c_3 - d_3 \\ A'_4 &= c_4 - d_4 \\ A'_5 &= c_5 - d_5 \\ A'_6 &= (c_6 - d_6) - 20\, d_3(c_3 - d_3) \\ &\text{etc.} \end{aligned}$$

To illustrate, we could select as a generating function

$$\theta_x = y_0 e^{-\frac{x^p}{q}}. \tag{6}$$

For $p = 2$ this becomes the normal curve of error, but θ_x as defined above would be more general than ϕ_x. Thus, one can readily select a value of p which will make A'_4 vanish for any distribution.

To illustrate again,

$$\theta_x = y_0 \operatorname{sech}^n ax \tag{7}$$

may be selected. This generating function has the same degree of freedom as (6) but possesses the added advantage of being integrable for integral values of n. Thus (7) does not necessitate the use of quadrature when dealing with a distribution of a continuous variable which is treated as such. (See Case V, p. 95.)

For skew distributions it is frequently desirable to choose a generating function which is essentially skew. Generating functions of this type which readily lend themselves to integration are rare.

If one chooses

$$\theta_x = y_0\left(1 + \frac{x}{a}\right)^m e^{-nx} \tag{8}$$

(Pearson's type III) as the generating function, the coefficient A_3 will automatically disappear if the third moments are equated. However, a value of a may be selected which will cause either A_4 or A_5 to vanish.

For a further study of generating functions which tend to make frequency series more rapidly convergent one may refer to *A Number of New Generating Functions with Application to Statistics*, a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Science in the University of Michigan, 1923, by Emeterio Roa.

Table IX, referred to above, is taken from this dissertation.

Corresponding to (5) it may be shown that any frequency distribu-

Table IX *a*. Graduation of Distribution of Heights of 6441 Colored Soldiers

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

Class Interval	Observed Frequencies	Graduated Results		
		One term of (3) used	Two terms of (3) used	Three terms of (3) used
(1)	(2)	(3)	(4)	(5)
146–147	0	0	0	1
148–149	2	2	1	2
150–151	9	6	3	5
152–153	13	13	10	12
154–155	23	29	24	26
156–157	56	59	54	53
158–159	88	110	106	102
160–161	162	188	189	180
162–163	318	292	300	289
164–165	468	419	434	425
166–167	564	551	570	569
168–169	665	665	682	691
170–171	708	737	746	762
172–173	749	752	749	767
174–175	747	700	685	697
176–177	586	599	580	582
178–179	469	471	453	447
180–181	314	340	328	317
182–183	207	225	221	211
184–185	133	137	139	132
186–187	70	77	81	78
188–189	38	39	44	45
190–191	22	19	23	24
192–193	15	8	11	13
194–195	10	3	5	6
196–197	3	1	2	3
198–199	2	0	1	1

TABLE IX *b*. TABLE SHOWING THE SUCCESSIVE STEPS IN GRADUATING THE DISTRIBUTION OF HEIGHTS OF 6441 COLORED SOLDIERS. GENERATING FUNCTION USED: $\phi(x) = Ke^{\frac{-x^2}{2\sigma^2}}$

BOUNDARY POINT OF CLASS INTERVAL		INTEGRAL OF GENERATING FUNCTION	INTEGRAL OF SECOND TERM	INTEGRAL OF THIRD TERM	INTEGRAL OF EXPANSION OF THREE TERMS	GRADUATED UNIT FREQUENCY THREE TERMS USED
x	$t=\frac{1}{\sigma}(x-b)$	$\int_{-\infty}^{t}\phi(t)dt$	$-\frac{C_3}{\underline{\lvert 3}}\phi^{(3)}(t)$	$\frac{C_4}{\underline{\lvert 4}}\phi^{(4)}(t)$	(3) + (4) + (5)	Δ (6)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
− 15.5	− 4.470	.00000	− .00001	.00001	.00000	.00002
− 14.5	− 4.176	.00001	− .00002	.00003	.00002	.00005
− 13.5	− 3.882	.00005	− .00006	.00008	.00007	.00015
− 12.5	− 3.587	.00017	− .00014	.00019	.00022	.00032
− 11.5	− 3.293	.00050	− .00033	.00037	.00054	.00080
− 10.5	− 2.999	.00135	− .00067	.00066	.00134	.00184
− 9.5	− 2.705	.00342	− .00123	.00099	.00318	.00399
− 8.5	− 2.411	.00795	− .00199	.00121	.00717	.00825
− 7.5	− 2.117	.01713	− .00280	.00109	.01542	.01577
− 6.5	− 1.823	.03415	− .00333	.00037	.03119	.02789
− 5.5	− 1.528	.06326	− .00314	− .00104	.05908	.04488
− 4.5	− 1.234	.10860	− .00185	− .00279	.10396	.06603
− 3.5	− .940	.17361	.00057	− .00419	.16999	.08830
− 2.5	− .646	.25914	.00358	− .00443	.25829	.10724
− 1.5	− .352	.36242	.00623	− .00312	.36553	.11829
− 0.5	− .058	.47687	.00752	− .00057	.48382	.11901
0.5	.237	.59367	.00694	.00222	.60283	.10827
1.5	.531	.70229	.00471	.00410	.71110	.09039
2.5	.825	.79531	.00172	.00446	.80149	.06934
3.5	1.119	.86843	− .00102	.00342	.87083	.04927
4.5	1.413	.92117	− .00278	.00171	.92010	.03273
5.5	1.707	.95609	− .00337	.00011	.95283	.02051
6.5	2.001	.97730	− .00307	− .00089	.97334	.01229
7.5	2.296	.98916	− .00231	− .00122	.98563	.00696
8.5	2.590	.99520	− .00151	− .00110	.99259	.00381
9.5	2.884	.99804	− .00086	− .00078	.99640	.00195
10.5	3.178	.99926	− .00044	− .00047	.99835	.00094
11.5	3.472	.99974	− .00020	− .00025	.99929	.00044
12.5	3.766	.99992	− .00008	− .00011	.99973	.00017
13.5	4.061	.99998	− .00003	−. 00005	.99990	.00006
14.5	4.355	.99999	− .00001	− .00002	.99996	.00003
15.5	4.649	1.00000	− .00000	− .00001	.99999	——

tion can be represented as a series proceeding by differences instead of derivatives, or

$$f_{(x)} = \frac{N}{N'}\left\{\theta_x - \frac{B_3}{\underline{|3}}\Delta^3\theta_x + \frac{B_4}{\underline{|4}}\Delta^4\theta_x - \frac{B_5}{\underline{|5}}\Delta^5\theta_x + \cdots\right\}. \qquad (9)$$

Charlier's type B curves are special cases of these series. Thus if the Poisson exponential binomial limit be selected,

$$\theta_x = \psi_x = \frac{e^{-\lambda}\lambda^x}{\underline{|x}}$$

we have Charlier's type B.

But since, for this generating function,

$$\nu_2 = \nu_3$$

and the third moment in practice is rarely as large as the second moment, a better selection of generating function with greater freedom, can usually be found. In general, the point binomial

$${}_nC_x p^{n-x} q^x$$

is better, since Poisson's binomial limit is merely a special case. The extra degree of freedom obtained by not restricting the generating function to a limiting value results in a better first approximation to the observed distribution and hence to a more rapidly converging series.

It must be borne in mind that any series involving derivatives or finite difference series can theoretically be used on a distribution of a continuous variable only by resorting to mechanical quadrature. (Case I, p. 92.) This applies to Charlier's type B series as well as to Pearson's generalized curves.

Practically, such series are, however, entirely satisfactory since the distribution may be treated as one of a discrete variable by leaving all moments unmodified and permitting the ordinates, instead of areas, to represent the graduated frequencies. (Case V, p. 95.)

CHAPTER VIII

SIMPLE CORRELATION

By H. L. RIETZ and A. R. CRATHORNE

MEANING OF CORRELATION[1]

Let us assume data consisting of pairs of corresponding values $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$ where we are interested in a quantitative description of the association of the x's and the corresponding y's. These values may arise from any one of a great variety of situations. For example, the x's may be the maximal daily temperatures of Boston and the y's the corresponding values for New York City; the x's may be statures of fathers and the y's those of their oldest sons; the x's may be the numbers of working hours per day of a group of laborers and the y's the corresponding wages of the group per day.

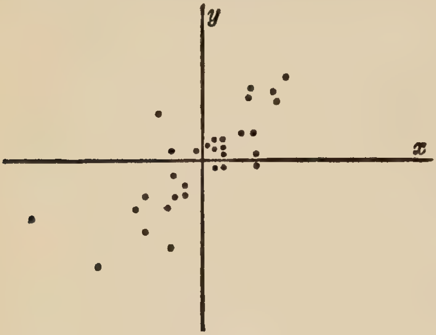


Fig. 7

Scatter-diagrams. If such a set of pairs of numbers is represented by a system of dots, marking the rectangular coördinates of points, we obtain a so-called "scatter-diagram." If we take the origin of coördinates so as to measure x's and y's from their respective mean values, the scatter-diagram of maximal daily temperatures of Boston and New York for July, 1920, is shown in Figure 7, where Boston temperatures are abscissas and New York temperatures are ordinates. It is clear from the plotting of such data that, with an assigned value of x, the corresponding value of y may have many values and thus cannot be accurately predicted by the use of a single-valued function of x. On the other hand, it is fairly obvious that for an assigned

[1] For definitions of correlation, see Karl Pearson, *Drapers' Company Research Memoirs, Biometric Series II* (1905), p. 9; H. L. Rietz, *Annals of Math.*, vol. 13 (1912), pp. 187–92; E. V. Huntington, *Amer. Math. Monthly*, vol. 26 (1919), pp. 423–27; A. R. Crathorne, *Report of National Committee on Mathematical Requirements* (1923), Chap. X, pp. 105–28.

value of x larger than the mean value of x's, a corresponding y taken at random is much more likely to be above than below the mean value of y. In other words, the x's and y's are not independent in the probability sense of independence. There is a tendency for the dots in Figure 7 to fall into a sort of band which can be fairly well described.

Briefly stated, there is an important field of association between perfect dependence given by a single-valued mathematical function at the one extreme and perfect independence in the probability sense at the other extreme. The theory of correlation is devoted to the description and characterization of this type of association.

THE CORRELATION COEFFICIENT

The correlation coefficient r. The most important measure of the degree of correlation is the so-called Pearsonian coefficient of correlation universally represented by the letter r. It is often called the product-moment coefficient.

Given the pairs of corresponding values

$$(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$$

of the variables X and Y measured from an arbitrary origin; let us take $\overline{X}$, $\overline{Y}$ for the arithmetic means of the given values of X's and Y's respectively. Then

$$x_i = X_i - \overline{X},$$
$$y_i = Y_i - \overline{Y}$$

are deviations from mean values, and the standard deviations[1] of the two series of values are

$$\sigma_x = \sqrt{\frac{1}{n}\Sigma(x_i^2)}, \qquad \sigma_y = \sqrt{\frac{1}{n}\Sigma(y_i^2)},$$

where the summation indicated by Σ extends from $i = 1$ to $i = n$.

Let the same values which are denoted by x in original units of the data (yards, pounds, kilograms, dollars) be denoted by x' when they are measured in the standard deviation σ_x as a unit. Similarly, let the values of y be denoted by y' when measured in σ_y as a unit. That is,

$$x_i' = \frac{x_i}{\sigma_x}, \qquad y_i' = \frac{y_i}{\sigma_y}.$$

Then, in terms of x_i' and y_i', the **correlation coefficient r** is given by the simple formula

$$r = \frac{1}{n}\Sigma x_i' y_i'. \tag{1}$$

[1] For the meaning of standard deviation see Chapter 2, p. 27.

That is, the correlation coefficient of two sets of values, expressed in their respective standard deviations as units, may be defined as the arithmetic mean of the products of deviations of corresponding values from their respective means.

Although the expression (1) for r is very simple for the purpose of giving a first notion of the meaning of r, the following formulas easily obtained from (1) are usually better adapted to numerical computation:

$$r = \frac{\frac{1}{n}\Sigma x_i y_i}{\sigma_x \sigma_y} = \frac{\Sigma x_i y_i}{\sqrt{\Sigma(x_i^2)}\,\sqrt{\Sigma(y_i^2)}} \tag{2}$$

$$= \frac{\frac{1}{n}\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_x \sigma_y} \tag{3}$$

$$= \frac{\frac{1}{n}\Sigma(X_i Y_i) - \overline{X}\,\overline{Y}}{\sigma_x \sigma_y} \tag{4}$$

$$= \frac{\frac{1}{n}\Sigma(X_i Y_i) - \overline{X}\,\overline{Y}}{\sqrt{\frac{1}{n}\Sigma(X_i^2) - \overline{X}^2}\;\sqrt{\frac{1}{n}\Sigma(Y_i^2) - \overline{Y}^2}} \tag{5}$$

Numerical computation of the correlation coefficient. For relatively small values of n, say $n < 30$, formula (5) is usually used where X and Y are the values given in the original[1] measurements.

If n is a large number, it usually saves labor to construct a so-called correlation table from the data as explained below and to organize the computations around the table.

In this connection, it is especially important to remember that the formulas (3), (4), and (5) hold when X and Y are measured from any arbitrary origin in any unit.

For dealing with large numbers, formula (4), with the values of X measured from a class mark near the mean of the X-series, and those of Y from a class mark near the mean of the Y-series, is usually most suitable.

Correlation table. A correlation table is simply a double-entry table constructed from the given data. In Table I is shown such a table from a ten years' record of the July maximal daily temperatures of Boston and New York. Such a table contains a system of columns and a system of rows each of which is a frequency distribution. The numbers

[1] J. A. Harris, *Amer. Naturalist*, vol. 44 (1910), pp. 693–99.

TABLE I. CORRELATION TABLE AND CALCULATION OF THE CORRELATION COEFFICIENT FOR BOSTON-NEW YORK MAXIMAL DAILY JULY TEMPERATURES FOR YEARS 1911–1920

X →
BOSTON

Y ↑
NEW YORK

	61	64	67	70	73	76	79	82	85	88	91	94	97	100	103	f_y	Y	$f_y \cdot Y$	$f_y \cdot Y^2$	T	$T \cdot Y$
97													1	2	1	4	+5	20	100	28	140
94							1		1		2	2	3	3	1	13	+4	52	208	67	268
91					2					1	3	1	3	1		11	+3	33	99	41	123
88					2	1	3	3	3	6	11	2				31	+2	62	124	76	152
85			1	1	1	4	7	9	4	8	5	3	2			45	+1	45	45	75	75
82			1	6	2	4	9	12	9	5	3					51	0			27	
79	1	1		5	5	7	22	8	10	6	1					66	−1	−66	66	7	−7
76		3	2	3	6	9	12	7	3							45	−2	−90	180	−40	80
73	1		5	5	3	3	5	3	1							26	−3	−78	234	−45	135
70		1	1	2	3	4										11	−4	−44	176	−25	100
67		2	1	2												5	−5	−25	125	−20	100
64	1			1												2	−6	−12	72	−9	54
																310		−103	1429		1220
f_x	3	7	11	25	24	32	59	42	31	26	25	8	9	6	2						
X	−6	−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5	+6	+7	+8						
$f_x \cdot X$	−18	−35	−44	−75	−48	−32		42	62	78	100	40	54	42	16	182					
$f_x \cdot X^2$	108	175	176	225	96	32		42	124	234	400	200	324	294	128	2558					
S	−10	−21	−27	−49	−27	−44	−44	−16	−5	17	43	18	28	25	9						
$S \cdot X$	60	105	108	147	54	44		−16	−10	51	172	90	168	175	72	1220					

$$\overline{Y} = \frac{-103}{310} = -0.3323$$

$$\overline{X} = \frac{182}{310} = 0.5871$$

$$\sigma_y^2 = \frac{1429}{310} - \overline{Y}^2 = 4.4993$$

$$\sigma_y = 2.1212$$

$$\sigma_x^2 = \frac{2558}{310} - \overline{X}^2 = 7.9069$$

$$\sigma_x = 2.8119$$

$$\frac{1}{n}\Sigma XY = \frac{1220}{310} = 3.9355$$

$$r = \frac{\frac{1}{n}\Sigma XY - \overline{X}\,\overline{Y}}{\sigma_x \sigma_y} = 0.6925$$

in a column corresponding to an assigned $X = X_1$ form what is often called an X array of Y's, and those in a row corresponding to $Y = Y_1$ a Y array of X's. In this special case, X represents Boston temperatures and Y, New York temperatures. Any number in a compartment of this table, say 10 in the column headed 85 and the row marked 79, is called a *cell* frequency and indicates that on 10 days of the total 310 days the maximal temperatures of Boston were between 83.5 and 86.5 and those of New York on the same days were between 77.5 and 80.5. A cell frequency of the s-th column and the t-th row will be denoted by n_{st}.

Form for the calculation of *r*. In calculating r it is important to have a systematic form in which to arrange the work to avoid confusion in the somewhat complicated details. Given the correlation table, we first add the frequencies in the rows and columns. This gives two total frequency distributions — the one with respect to temperatures of Boston in the row marked f_x and the other with respect to the temperatures of New York in the column marked f_y. Arbitrary origins are taken near the means and the class intervals chosen as units of measurement. This gives the column headed Y and the row marked X. The next two columns and the next two rows are self-explanatory and are used in calculating the means $\bar{X}$ and $\bar{Y}$, and the standard deviations σ_x, σ_y by the method explained on page 28. We find

$\bar{X} = 0.5871$ in class intervals as the unit, and measured from 79°.
$\sigma_x = 2.8119$ in class intervals as the unit.
$\bar{Y} = -0.3323$ in class intervals as the unit, and measured from 82°.
$\sigma_y = 2.1212$ in class intervals as the unit.

As for the next column, the heading T is an abbreviation for $\sum_{a_t} n_{st}X$, which means the sum of the products of each cell frequency of the array t (row) and the corresponding value of X. Thus to find the fourth number, 76, we have

$$\begin{array}{rcrcr}
(-2) & \times & 2 & = & -4 \\
(-1) & \times & 1 & = & -1 \\
0 & \times & 3 & = & 0 \\
1 & \times & 3 & = & 3 \\
2 & \times & 3 & = & 6 \\
3 & \times & 6 & = & 18 \\
4 & \times & 11 & = & 44 \\
5 & \times & 2 & = & \underline{10} \\
 & & & & 76 = T \text{ for the fourth row.}
\end{array}$$

In a similar way the S marking the next to last row is an abbreviation of $\sum_{a_s} n_{st} Y$, which means the sum of the products of the cell frequencies of the array s (column) and the corresponding value of Y. Thus, to find the third number in the rows marked S, we have

$$\begin{array}{rcr}
(+1) \times 1 & = & +\ \ 1 \\
0 \times 1 & = & 0 \\
-1 \times 0 & = & 0 \\
-2 \times 2 & = & -\ \ 4 \\
-3 \times 5 & = & -15 \\
-4 \times 1 & = & -\ \ 4 \\
-5 \times 1 & = & \underline{-\ \ 5} \\
 & & -27
\end{array} = S \text{ for the third column.}$$

The column headed $T \cdot Y$ is formed in an obvious manner by multiplying together the corresponding numbers in the T column and the Y column. The lower row marked $S \cdot X$ is formed in a similar manner from the products of corresponding numbers in the S and the X rows. It is evident that the total of the last column or of the last row is $\Sigma(XY)$ ($= 1220$) for all the 310 entries in the table and hence we have a good check on the accuracy of the computation.

We have now computed $\overline{X}$, $\overline{Y}$, $\Sigma(XY)$, σ_x and σ_y and substitution in formula (4) gives for the value of the coefficient of correlation

$$r = 0.6925.$$

Probable error of r. If n is a fairly large[1] number and if the regression is not far from linear (see p. 126), the probable error of r is given by the formula

$$PE = 0.6745 \frac{1 - r^2}{\sqrt{n}} \cdot$$

When applied to the above calculation of r, we have $r = 0.6925 \pm 0.0199$, or $r = 0.693 \pm 0.020$ if we use three places of decimals.

Other formulas for r. It is easily verified by comparison with (3) that[2]

$$r = 1 - \frac{1}{2\,n} \Sigma(x'_i - y'_i)^2 \tag{6}$$

$$= -1 + \frac{1}{2\,n} \Sigma(x'_i + y'_i)^2 \tag{7}$$

$$= 1 - \frac{\sigma^2_{(x'-y')}}{2} = -1 + \frac{\sigma^2_{(x'+y')}}{2}, \tag{8}$$

[1] For errors of sampling with small numbers, see Student, *Biometrika*, vol. 6 (1908–9), pp. 302–10. Also Soper, Young, Cave, Lee, and Pearson, *Biometrika*, vol. 11 (1915–17), pp. 328–413.

[2] See Huntington, *Amer. Math. Monthly*, vol. 26 (1919), p. 424.

where $\sigma_{(x'-y')}$ and $\sigma_{(x'+y')}$ are the standard deviations of $(x' - y')$ and $(x' + y')$ respectively.

From these formulas, we have at once the important property that r is not less than -1 nor greater than $+1$.

It is easily verified by comparison with formula (2) that

$$r = \frac{\sigma_x{}^2 + \sigma_y{}^2 - \sigma^2_{(x-y)}}{2\,\sigma_x\sigma_y}. \tag{9}$$

When n is a small number, the values of r may in certain cases be obtained very simply from (9).

REGRESSION

Lines of regression. If we mark the mean temperature for each column of the "scatter-diagram" (Fig. 8) by a cross, the "line of regression of Y on X" is the straight line

$$Y = mX + b, \tag{10}$$

which fits "best"[1] this system of crosses.

The line of regression of Y on X has the property that the sum of the squares of the distances (measured parallel to the Y-axis) of all the dots is less than from any other straight line. Moreover, the values of Y computed from the regression line are more highly correlated with the corresponding observed Y's than when calculated from any other linear function of X.

For this reason, the equation of the line of regression of Y on X may be regarded as that linear relation which on the whole gives the "best" estimate of Y corresponding to an assigned X in so far as such a prediction can be made by means of a linear function.

It is easily shown that (10) takes the form

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X}), \tag{11}$$

or

$$y = r\frac{\sigma_y}{\sigma_x}x, \text{ where } y = Y - \overline{Y} \text{ and } x = X - \overline{X},$$

or

$$y' = rx', \text{ where } x' = \frac{x}{\sigma_x} \text{ and } y' = \frac{y}{\sigma_y}.$$

[1] The term "best" is here used to mean the best under a least-squares criterion of approximation. In applying the criterion, the squares of distances of the crosses from the line are weighted with the number of dots in the corresponding column. See G. Udny Yule, *Proc. Roy. Soc.*, vol. 60 (1897), pp. 477–89; *Introduction to the Theory of Statistics* (1915), pp. 168–75.

Thus, given X the maximal daily temperature of Boston, we have for the best estimate of the maximal daily temperature Y of New York

$$Y = \overline{Y} + r\frac{\sigma_y}{\sigma_x}(X - \overline{X}) = 81.00 + .5224(X - 80.76) \qquad (12)$$
$$= .5224X + 38.81,$$

in so far as such a prediction can be made by means of a linear equation. This tells us that corresponding to an assigned change of 1 degree in X, there is on the average a change of .5224 degrees in Y.

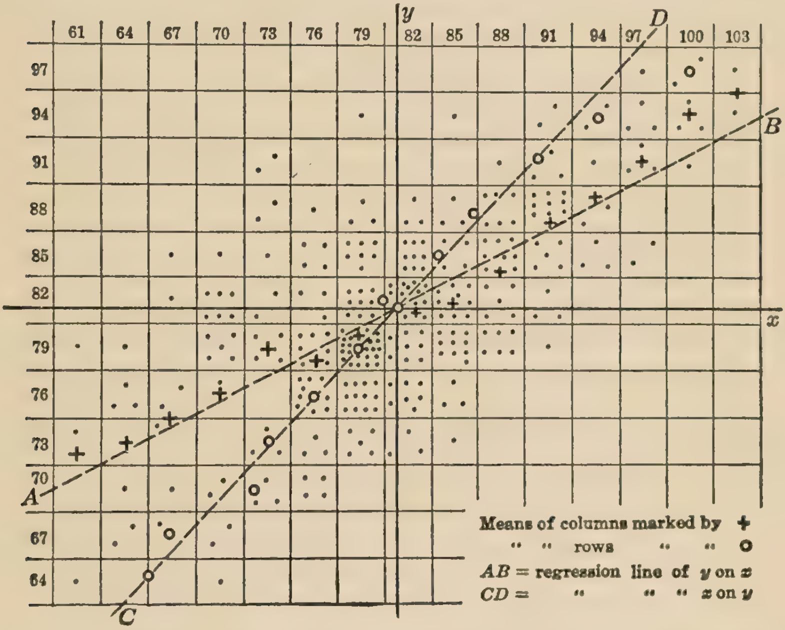


FIG. 8. SCATTER-DIAGRAM FOR DATA OF TABLE I, SHOWING CENTER OF TABLES, MEANS OF ARRAYS, AND LINES OF REGRESSION

In a similar manner, we find the line of regression of X on Y to be

$$X = \overline{X} + r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y}) = .9180\,Y + 6.40. \qquad (13)$$

It is important to note that the value of X in (13) cannot be obtained by solving for X in (12).

When there is no correlation between X's and Y's, $r = 0$; and except for chance fluctuations, (12) and (13) are parallel to the X and Y axes respectively. But conversely, when $r = 0$ it is not necessarily true that there is no correlation. There may be high correlation[1] with non-linear regression when $r = 0$.

The mean square error of estimate — standard deviation of arrays. In estimating Y as just shown from the regression equation of Y on X, it is important to know something about the variability in the Y arrays. The mean square error in estimating Y's, by taking the means of arrays of Y's may be defined as the mean of the squares of the standard deviation of these arrays, the square of each standard deviation of an array being weighted with the number in the array. If the means of these arrays fall exactly on the line of regression of Y's on X's, then it can be proved that

$$s_y^2 = \sigma_y^2(1 - r^2), \tag{14}$$

where s_y^2 is the mean square error in the estimate of Y. The value s_y^2 may also be defined as the mean of the squares of the deviations of the dots of the scatter-diagram from the line of regression of Y on X, when distances are measured parallel to the Y-axis.

From (14) we have

$$s_y = \sigma_y\sqrt{1 - r^2}. \tag{15}$$

This value of s_y may be regarded as a sort of average value of the standard deviations of the arrays of Y's and is sometimes called the root mean square error of estimate of Y, or more briefly, the standard error of estimate of Y. The factor $\sqrt{1 - r^2}$ in (15) has been called the coefficient of **alienation** or the measure of the failure to improve the estimate of Y from knowledge of the correlation.

To illustrate, suppose that $r = .6925$ be the correlation coefficient of maximal daily temperatures of Boston and New York. Assuming linear regression the root mean square error of estimate of the temperature of New York from that of Boston would be

$$s_y = \sigma_y\sqrt{1 - r^2} = 0.7214\,\sigma_y. \tag{16}$$

Hence the average variability in the arrays of New York temperatures which correspond to assigned Boston temperatures is more than .7 as great as the average variability of all the New York temperatures. We find therefore that we cannot, with any considerable degree of reliability, predict on a given day the maximal temperature of New York from that of Boston. However, with large numbers, we can give a very reliable prediction of the mean maximal daily New York temperature that corresponds to an assigned Boston temperature.

[1] H. L. Rietz, "On Functional Relations for which the Coefficient of Correlation is Zero," *Quar. Pub. Amer. Stat. Assoc.*, vol. 16 (1919), pp. 472–76.

An analogous discussion of the estimation of X from the means of the arrays of X's leads to the equation

$$s_x = \sigma_x\sqrt{1 - r^2}.$$

THE CORRELATION RATIO

Non-linear regression.[1] **The correlation ratio.** When the means of a system of parallel arrays do not lie near a straight line, the regression is said to be non-linear. We then speak of the curve formed by the means of arrays of Y's as the regression curve of Y on X. In the case of linear regression we found, (14),

$$s_y^2 = \sigma_y^2(1 - r^2),$$

or

$$r^2 = 1 - \frac{s_y^2}{\sigma_y^2}. \tag{17}$$

If we think of s_y as measured in terms of σ_y as a unit, the discrepancy of r^2 from unity, or the departure from perfect correlation is measured by s_y^2. When s_y approaches σ_y, then $1 - \frac{s_y^2}{\sigma_y^2}$ approaches zero. If s_y^2 is small, that is, if the points of the scatter-diagram tend to concentrate in a narrow band the expression $1 - \frac{s_y^2}{\sigma_y^2}$ approaches unity. This suggests the use of $1 - \frac{s_y^2}{\sigma_y^2}$ as a measure of correlation for linear or for non-linear regression. We then write

$$\eta_{yx}^2 = 1 - \frac{s_y^2}{\sigma_y^2}, \tag{18}$$

where η_{yx} is called the **correlation ratio**[2] of y on x. For linear regression, we have $\eta = r$.

An analogous discussion for the arrays of X's leads to the equation

$$\eta_{xy}^2 = 1 - \frac{s_x^2}{\sigma_x^2}, \tag{19}$$

giving η_{xy}, the correlation ratio of x on y.

In general we may say that the correlation ratio is a measure of the concentration of the dots of the scatter-diagram about the regression curve.

[1] Karl Pearson, "On the General Theory of Skew-Correlation and Non-linear Regression," *Drapers' Company Research Memoirs, Biometric Series II* (1905).

[2] Karl Pearson, *loc. cit.*, p. 10.

By a slight transformation, (18) and (19) can be put into the forms

$$\eta_{yx}{}^2 = \frac{\sigma_{m_y}{}^2}{\sigma_y{}^2},\ \eta_{xy}{}^2 = \frac{\sigma_{m_x}{}^2}{\sigma_x{}^2}, \tag{20}$$

where σ_{m_y}, σ_{m_x} are the standard deviations[1] of the means of the X arrays of Y's and of the Y arrays of X's respectively. The expression "correlation ratio" probably had its origin in the ratios given in (20).

Form for the calculation of correlation ratios. Referring to the notation used in the computation of the correlation coefficient it can easily be shown[2] that the two correlation ratios can be expressed in the form

$$\eta_{yx}{}^2 = \left\{\frac{1}{n}\sum\left(\frac{S^2}{f_x}\right) - \bar{Y}^2\right\}\frac{1}{\sigma_y{}^2}, \tag{21}$$

$$\eta_{xy}{}^2 = \left\{\frac{1}{n}\sum\left(\frac{T^2}{f_y}\right) - \bar{X}^2\right\}\frac{1}{\sigma_x{}^2}. \tag{22}$$

To compute the two ratios in connection with the correlation coefficient, it is then necessary to add two columns and two rows to the form, p. 123, for the computation of r. For example, in the temperature problem the two columns and two rows are given below:

T^2	$T^2 \div f_y$	S^2	$S^2 \div f_x$
784	196.00	100	33.33
4489	345.31	441	63.00
1681	152.82	729	66.27
5776	186.32	2401	96.04
5625	125.00	729	30.38
729	14.29	1936	60.50
49	0.74	1936	32.81
1600	35.56	256	6.10
2025	77.88	25	0.81
625	56.82	289	11.12
400	80.00	1849	73.96
81	40.50	324	40.50
		784	87.11
		625	104.17
		81	40.50

$$\frac{1}{n}\sum\left(\frac{T^2}{f_y}\right) = \frac{1311.24}{310}$$

$$\frac{1}{n}\sum\left(\frac{S^2}{f_x}\right) = \frac{746.60}{310}$$

Having summed up the column marked $T^2 \div f_y$ and the row marked

[1] In finding these standard deviations, the squares of the deviations of the means of arrays from the mean of all values are weighted with the numbers in the arrays.

[2] A. R. Crathorne, "Calculation of the Correlation-Ratio," *Quar. Pub. Amer. Stat. Assoc.*, vol. 18 (1922), pp. 394–96.

$S^2 \div f_x$, the values of the two correlation ratios are then easily computed from (21) and (22), giving

$$\eta_{yx} = 0.7146, \quad \eta_{xy} = 0.7010.$$

Probable error of η. The probable error of a correlation ratio is given approximately by the expression

$$PE = \frac{0.6745}{\sqrt{n}}(1 - \eta^2). \tag{23}$$

When applied to the above calculations of η_{yx} and η_{xy}, we have

$$\eta_{yx} = 0.7146 \pm 0.0187, \quad \eta_{xy} = 0.7010 \pm 0.0195.$$

Test of linearity of regression. Since r is a good measure of correlation only if we have nearly linear regression, it is necessary to examine our data for linearity of regression. A necessary and sufficient condition for linearity is that $\eta^2 - r^2$ shall differ from zero by an amount not greater than the fluctuations due to random sampling. A common test[1] is the comparison of this difference with its probable error,

$$PE = 0.6745 \frac{2}{\sqrt{n}} \sqrt{(\eta^2 - r^2)\{(1 - \eta^2)^2 - (1 - r^2)^2 + 1\}}. \tag{24}$$

If $\eta^2 - r^2$ is small compared with r, or if η and r are both small, an easily calculated test arising out of (24) is

$$\frac{\sqrt{n}}{0.6745} \cdot \tfrac{1}{2}\sqrt{\eta^2 - r^2} < 2.5, \tag{25}$$

or

$$n(\eta^2 - r^2) < 11.37. \tag{26}$$

For our temperature problem we find

$$n(\eta_{yx}{}^2 - r^2) = 9.64, \quad n(\eta_{xy}{}^2 - r^2) = 3.66,$$

and the test for linearity is satisfied for both regression lines.

Corrections for grouping and errors of observation. In using certain class intervals for grouping data in finding the correlation coefficient or the correlation ratio, it is important to correct the computed values for grouping. Sometimes it is also desirable to correct for errors of observation. Certain appropriate corrections have been published.[2]

[1] J. Blakeman, "On Tests for Linearity of Regression," *Biometrika*, vol. 4 (1906), pp. 332–50.

[2] G. Udny Yule, *Introduction to the Theory of Statistics* (Edition 1922), pp. 211–14.

T. L. Kelley, *Statistical Method* (1923), p. 168.

Karl Pearson, "On the Correction to be Made in the Correlation Ratio," *Biometrika*, vol. 8 (1911–12), pp. 254–56.

Student, "The Correction to be Made in the Correlation Ratio for Grouping," *Biometrika*, vol. 9 (1913), pp. 316–20.

FURTHER METHODS OF DETERMINING CORRELATION

Correlation from ranks. When data are not measurements expressed in cardinal numbers, but are merely marks of the orders or ranks of individuals in a series, we may seek a measure of correlation between corresponding ranks. For example, the following table gives the ranks of ten students in two tests:

Student	A	B	C	D	E	F	G	H	I	J
Rank in first test	1	2	3	4	5	6	7	8	9	10
Rank in second test	2	4	1	3	6	5	7	9	10	8
Changes in rank	1	2	2	1	1	1	0	1	1	2

Let ν_x and ν_y be the ranks of corresponding variables in two series of n individuals each, then the correlation of ν_x and ν_y is given by [1]

$$\rho = 1 - \frac{6\,\Sigma\,(\nu_x - \nu_y)^2}{n(n^2 - 1)}. \qquad (27)$$

Applying this to the student tests above, we have

$$\rho = 0.891 \pm 0.047,$$

where the probable error is obtained from the expression

$$PE = \frac{0.6745(1 - \rho^2)}{\sqrt{n}}\,\{1 + .086\,\rho^2 + .013\,\rho^4 + .002\,\rho^6\}. \qquad (28)$$

Under the assumption of a normal frequency distribution, the corresponding value of the correlation coefficient of the variables that correspond to these ranks is given by

$$r = 2 \sin\left(\frac{\pi}{6}\,\rho\right) \qquad (29)$$

with a probable error

$$PE = \frac{0.7063(1 - r^2)}{\sqrt{n}}\,\{1 + .042\,r^2 + .008\,r^4 + .002\,r^6\}. \qquad (30)$$

For the student rank data we find

$$r = .900 \pm .043.$$

It is easily shown that the correlation between actually measured variables can be made to change very much without changing ranks. Thus the two series

60, 50, 40, 30, 20
100, 99, 98, 3, 1

[1] Karl Pearson, "On Further Methods of Determining Correlation," *Drapers' Company Research Memoirs, Biometric Series* IV (1907), p. 13.

are illustrations of perfect correspondence in rank, but the correlation of the actual numbers is far from perfect. The nature of the distribution is a fundamental consideration in the case of correlation in ranks.

Ties in rank. The question of appropriate ranking arises when two or more values are equal. In this case two methods are in use:

(*a*) **The bracket rank method.** In this method the individual values of the ties are assigned the same rank and one greater than that of the individual which immediately preceded the ties. The next individual after the ties takes the rank it would have had if each of the preceding individuals in the ties had different ranks.

(*b*) **The mid-rank method.** In this method, the individual values of the tie are given equal rank but that rank is the value at the middle of the set of ties. The following illustrates both methods:

MEASUREMENT	BRACKET METHOD	MID-RANK METHOD
75	1	1
80	2	2
85	3	4
85	3	4
85	3	4
90	6	6.5
90	6	6.5
95	8	7

Spearman's foot-rule. Spearman has suggested[1] a very simple formula for finding correlation from ranks which he called the "foot-rule" formula,

$$R = 1 - \frac{L}{M}, \tag{31}$$

where L denotes the sum of the positive differences in rank and $M = \frac{n^2 - 1}{6}\cdot$ As its name indicates, this is a very rough estimate of the correlation between the variables. When we have a normal distribution of the actual measurements, R does not in general approximate r. On the assumption of a normal distribution Pearson has shown that

$$r = 2 \cos \frac{\pi}{3}(1 - R) - 1. \tag{32}$$

Tetrachoric correlation. The problem of measuring correlation sometimes arises in connection with variables which do not admit of exact measurement, or only admit of it with very great labor; for example, colors, temperaments, shapes, and so on. Nevertheless, it may

[1] C. Spearman, "A Foot-rule for Measuring Correlation," *Brit. Journal of Psychology*, vol. 2 (1906), p. 89; also vol. 3 (1910), p. 271.

appear legitimate to assume for purposes of determining correlation, that there lie back of our categorical classifications of these variables certain quantitative values that are normally distributed. Again we may have measurable variables in very broad categories, for example, high and low temperatures; tall, medium, and short individuals; grades A, B, C, D, and E in school subjects. If there are only two categories, we get a fourfold table of the form

a	b
c	d

Such a table would arise in our Boston-New York temperatures problem if we simply recorded high or low maximal temperatures. If we consider all temperatures above 77.5 as high and other temperatures low, our table of page 123 would reduce to

		Boston	
		High	Low
New York	High	177	44
	Low	31	58

Under the assumption of a distribution conforming to the normal frequency surface, Pearson has shown that the correlation coefficient r (called **tetrachoric** r) is given by a root of the equation

$$\frac{d}{n} = \tau_0\tau'_0 + \tau_1\tau'_1 r + \tau_2\tau'_2 r^2 + \tau_3\tau'_3 r^3 + \tau_4\tau'_4 r^4 + \tau_5\tau'_5 r^5 + \tau_6\tau'_6 r^6 + \cdots,$$

where τ_0[1] $= \dfrac{b+d}{n}$, $\tau'_0 = \dfrac{c+d}{n}$, and the other τ's, called tetrachoric functions, are given to τ_6 in *Tables for Statisticians and Biometricians*, pp. 42÷51. Methods are shown on page l of these tables for obtaining values of τ_m when $m \geqq 7$, if they are needed.

For the above illustration the equation of degree six in r is

$$.18710 = .09446 + .12323\,r + .01532\,r^2 + .01130\,r^3 + .00961\,r^4 + .00231\,r^5 + .00665\,r^6,$$

from which we find $r = .653$.

If the division between high and low temperatures had been 80.5 degrees, we would find $r = .663$. If the data were distributed normally, the value of r would not be changed by changing the point dividing high temperatures from low temperatures. The tetrachoric r for a

[1] The separation into a fourfold table should be such that $\tau_0 \leqq .5$ and $\tau'_0 \leqq .5$.

normal distribution agrees with the correlation coefficient for the case of an indefinitely fine grouping of measurable values.

The probable error [1] of tetrachoric r is relatively much larger than the probable error of r calculated from measurements — often in the neighborhood of three times the probable error of the r calculated from measurements.

The Yule coefficients. As a method of measuring association between attributes — say between vaccination and recovery from smallpox — Yule devised a " coefficient of association " [2] for a fourfold table which is designated by the letter "Q," where

$$Q = \frac{ad - bc}{ad + bc}. \tag{34}$$

In comparison with other measures of association, this coefficient seems to have no advantages except its simplicity.

In 1912, Yule published [3] another coefficient

$$\omega = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}, \tag{35}$$

which he termed the " coefficient of colligation."

Yule has expressed his preference for ω rather than Q as a measure of association. The use of both these coefficients has been the subject of vigorous criticism by Pearson and Heron.[4]

Method of contingency.[5] In this method, the degree of contingency between two variables is measured by a function of the difference between the numbers actually found in the cells of a correlation table and the numbers that would be found if the two variables were independent in the probability sense.

In the notation of page 124, the mean square contingency is defined by

$$\phi^2 = \frac{1}{n}\sum\left\{\frac{\left(n_{st} - \frac{f_s f_t}{n}\right)^2}{\frac{f_s f_t}{n}}\right\}, \tag{36}$$

the summation to extend to all compartments of the table.

[1] For the method of calculation of the probable error of tetrachoric r, see *Tables for Statisticians*, etc. (1914), pp. xl–xlii; cf. *Biometrika*, vol. 9 (1913), pp. 22–27.

[2] *Phil. Trans. Roy. Soc.*, Series A, vol. 194 (1900), p. 257; G. Udny Yule, *Introduction to the Theory of Statistics* (1915), p. 38.

[3] *Journal of the Royal Statistical Soc.*, vol. 75 (1911–12), pp. 579–652.

[4] "On the Theories of Association," *Biometrika*, vol. 9 (1913), pp. 159–315.

[5] Karl Pearson, "On the Theory of Contingency and its Relation to Association and Normal Correlation," *Drapers' Company Research Memoirs, Biometric Series I* (1904).

The coefficient of mean square contingency is defined by

$$C_1 = \sqrt{\frac{\phi^2}{1 + \phi^2}}. \tag{37}$$

If temperatures above 86.5 are called high, those below 74.5 low, and the others medium, the Boston-New York table of temperatures, page 123, becomes:

		Boston			
		Low	Medium	High	f_t
New York	High	4	12	43	59
	Medium	38	136	33	207
	Low	28	16	0	44
	f_s	70	164	76	

From this we find $\phi^2 = .4349$ and $C_1 = .551$.

With normal distribution and certain restrictions as to grouping, C_1 becomes numerically equal to r.

The method of contingency is chiefly used in measuring correlation between variables not capable of quantitative measurement, though there may be many graduations. One objection to the use of C_1 is that it varies with the grouping in the table. This, however, may be overcome by certain corrections.[1] When the data are grouped in κ rows and λ columns, and are to be considered as a random sample, the chief correction for grouping is the subtraction of

$$\frac{(\kappa - 1)(\lambda - 1)}{n}$$

from ϕ^2, C_1 being then calculated from the corrected ϕ^2.

Bi-serial *r*. We wish sometimes to find the correlation between two variables, one of which is measurable while the other is given only in alternative categories.[2] For example, X may be the grade of a student in mathematics while Y is simply " athletic " or " non-athletic "; or X may record the temperatures of Boston in many groups while Y, the temperature of New York, may be recorded merely as either high or low, giving a two-rowed correlation table as shown in the diagram.

[1] Karl Pearson, "On the Measurement of the Influence of 'Broad Categories' on Correlation," *Biometrika*, vol. 9 (1913), pp. 116–39 and 216–17.

[2] Karl Pearson, "On a New Method of Determining Correlation between a Measured Character A, and a Character B of which only the Percentage of Cases wherein B exceeds (or falls short of) a given intensity is recorded for each Grade of A," *Biometrika*, vol. 7 (1909–10), pp. 96–105.

TABLE II. BI-SERIAL CORRELATION TABLE FOR BOSTON-NEW YORK MAXIMAL DAILY JULY TEMPERATURES FOR YEARS 1911–1920

	$X \rightarrow$							BOSTON								
NEW YORK $Y \uparrow$		61	64	67	70	73	76	79	82	85	88	91	94	97	100	103	
	High			1	1	5	5	11	12	8	15	21	8	9	6	2	104
	Low	3	7	10	24	19	27	48	30	23	11	4					206
																	310

If we assume linear regression of the X on Y and that the distribution of the variable Y given in alternative groupings is approximately normal, then the correlation coefficient, called bi-serial r, is found as follows:

Let $\overline{X}_s$ be the mean of the row containing the smaller number, n_s, of entries, $\overline{X}$ the mean of all the X's, σ_x the standard deviation of the X's, and n the total number of entries in the table. Bi-serial r is then found by dividing

$$\frac{n_s}{n} \cdot \frac{\overline{X}_s - \overline{X}}{\sigma_x}$$

by the value of $\phi(t)$ in the tables on page 209 corresponding to $\frac{1}{2} - \frac{n_s}{n}$ in the column headed $\int_0^t \phi(t)dt$.

In the above table of temperatures, we find $\overline{X} = 80.762$, $X_s = 87.279$, $\sigma_x = 8.4357$, $n = 310$, $n_s = 104$, $\frac{n_s}{n} = .3355$, from which we obtain

$$\frac{n_s}{n} \cdot \frac{\overline{X}_s - \overline{X}}{\sigma_x} = .2592.$$

From table, p. 209, corresponding to $\frac{1}{2} - \frac{n_s}{n} = .1645$ in the $\int_0^t \phi(t)dt$ column we find $\phi(t) = .3645$, hence bi-serial r is $\frac{.2592}{.3645} = .711$.

The probable error of bi-serial r is greater than the probable error for the product moment coefficient.[2] Under certain distributions it may be twice as great.

Pearson has given a method for finding the correlation when X is given in multiple categories and Y in alternative categories, no assumptions being made as to regression or distribution.[3]

[1] H. E. Soper, "On the Probable Error of the Bi-serial Expression for the Correlation Coefficient," *Biometrika*, vol. 10 (1914–15), pp. 384–90.

[2] Karl Pearson, "On a New Method of Determining Correlation when one Variable is Given by Alternative and the Other by Multiple Categories," *Biometrika*, vol. 7 (1909–10), pp. 248–57.

Difficulties in the interpretation of correlation. Although the theory of correlation is very useful in the quantitative description and characterization of phenomena, many difficulties arise in the interpretation of the correlation coefficient. For example, the values of two variables x and y may be uncorrelated with each other and with a third variable z, but the quotients $\frac{x}{z}$ and $\frac{y}{z}$, or the products xz, yz would show decided correlation.[1]

Again, two variables may be uncorrelated in each of two records, but show correlation in mixed records.[2] For example, take one set of records consisting of n pairs of uncorrelated items with mean values $(\bar{x}, \bar{y})$, and another set of n pairs of uncorrelated items with mean values $(\bar{x}', \bar{y}')$, where $\bar{x}'$ and $\bar{y}'$ are respectively greater than $\bar{x}$ and $\bar{y}$. Even with the zero correlations in each of the two sets of n, we may nevertheless have a considerable positive correlation coefficient r from the $2n$ pairs of values obtained by combining the two sets.

In some cases part or all of the correlation between two variables may be traceable to a third variable whose effect could be eliminated by the methods of partial correlation theory. (See Chap. IX.)

[1] Karl Pearson, "On a Form of Spurious Correlation that may Arise when Indices are Used in the Measurement of Organs," *Proc. Roy. Soc.*, vol. 60 (1897), p. 489.

G. Udny Yule, "On the Interpretation of Correlations between Indices or Ratios," *Jour. Royal Stat. Soc.*, vol. 73 (1910), p. 644.

[2] Karl Pearson, *Phil. Trans.*, vol. 192 (1899), p. 257; cf. G. Udny Yule, *Introduction to the Theory of Statistics* (1917), pp. 218–19.

CHAPTER IX

PARTIAL AND MULTIPLE CORRELATION

By TRUMAN L. KELLEY

GENERALIZED CORRELATION COEFFICIENTS

Meaning, notation, and formulas. The basic problem of multiple correlation is to estimate the value of a variable that corresponds to assigned values of two or more other variables. For example, we may have assigned values of the maximal daily temperatures of Boston, Philadelphia, and Buffalo, and seek from these values the best estimate of the temperature of New York City. The use of the linear regression equation in two variables for estimating Y when X is assigned is presented on page 126. It may be recalled that the mathematical advantage of linear regression in dealing with two variables is that it leads to a simple equation permitting the estimation of one variable knowing the other.

We now seek an appropriate extension of the method of linear regression to n variables $X_1, X_2, \cdots X_n$. To be more precise, let us assume that we may estimate X_1 from assigned values of $X_2, X_3, \cdots X_n$ by means of the linear function

$$\overline{X}_1 = b_{12.34\ldots n} X_2 + b_{13.24\ldots n} X_3 + \cdots + b_{1n.23\ldots n-1} X_n + c, \qquad (1)$$

in which the b's and the c are constants, so chosen that the X_1's computed from (1) are the " best " (the meaning of ".best " as here used is given in Chapter 8, page 126) estimates of the observed X_1's which can be made by means of a linear function of the assigned values of $X_2, X_3, \cdots X_n$. The ordinary product-moment coefficient of correlation between the X_1 as thus estimated and the observed X_1 is called the " multiple " correlation coefficient.

If measures are expressed as deviations from their own means and divided by their own standard deviations, the regression equation (1) simplifies. Let

$$\bar{z}_1 = \frac{\overline{X}_1 - M_1}{\sigma_1}, \quad z_2 = \frac{X_2 - M_2}{\sigma_2}, \text{ etc.,}$$

in which the M's are successive means and the σ's successive standard deviations. The regression equation connecting the z's is

$$\bar{z}_1 = \beta_{12.34 \cdots n} z_2 + \beta_{13.24 \cdots n} z_3 + \cdots + \beta_{1n.23 \cdots n-1} z_n, \tag{2}$$

in which the b's and β's are connected by the relationships:

$$\begin{aligned} b_{12.34 \cdots n} &= \beta_{12.34 \cdots n} \frac{\sigma_1}{\sigma_2}, \\ b_{13.24 \cdots n} &= \beta_{13.24 \cdots n} \frac{\sigma_1}{\sigma_3}, \text{ etc.} \end{aligned} \tag{3}$$

Also:

$$c = M_1 - b_{12.34 \cdots n} M_2 - b_{13.24 \cdots n} M_3 - \cdots - b_{1n.23 \cdots n-1} M_n. \tag{4}$$

Knowing equation (2) and having relationships (3) and (4), it is but a step to secure equation (1), which is the most serviceable form of the regression equation for actual use.

The problem is then the determination of the values of the β constants in terms of the total correlation coefficients. If z_1 is the value of the first variable and $\bar{z}_1$ the estimated value as defined in (2), then $(z_1 - \bar{z}_1)$ is an error of estimate. Note that $\Sigma z_1 = 0$, and inspection of (2) shows that $\Sigma \bar{z}_1 = 0$; therefore $(z_1 - \bar{z}_1)$ is a deviation from a mean, and the standard deviation of such errors of estimate, which we will represent by the symbol $k_{1.23 \cdots n}$, is the standard error of estimate of z_1 variates. It is given by

$$k^2_{1.23 \cdots n} = \frac{\Sigma(z_1 - \bar{z}_1)^2}{N}.$$

Further,

$$\sigma_1^2 k^2_{1.23 \cdots n} = \frac{\Sigma(\sigma_1 z_1 - \sigma_1 \bar{z}_1)^2}{N} = \frac{\Sigma(x_1 - \bar{x}_1)^2}{N} = \sigma^2_{1.23 \cdots n},$$

so that $\sigma_{1.23 \cdots n}$, the standard error[1] of estimate of X_1 variates, is given by

$$\sigma_{1.23 \cdots n} = \sigma_1 k_{1.23 \cdots n}. \tag{5}$$

If $r_{1.23 \cdots n}$ is the correlation between X_1 and the linear function of $X_2, X_3, \cdots X_n$ from which we estimate X_1, then by parity with formula (14) of Chapter 8, page 128 $(\sigma_{1.2} = \sigma_1\sqrt{1 - r^2_{12}})$ we have

$$\sigma_{1.23 \cdots n} = \sigma_1 \sqrt{1 - r^2_{1.23 \cdots n}}, \tag{6}$$

so that

$$k_{1.23 \cdots n} = \sqrt{1 - r^2_{1.23 \cdots n}}, \text{ or, } r_{1.23 \cdots n} = \sqrt{1 - k^2_{1.23 \cdots n}}. \tag{7}$$

The reader will note that since $r^2_{1.23 \cdots n} + k^2_{1.23 \cdots n} = 1$ the correlation and alienation coefficients are related to each other as are the sine and cosine of an angle.

[1] *Cf.* p. 128.

It only remains to find $k_{1.23\ldots n}$, the multiple alienation coefficient, in order to determine $r_{1.23\ldots n}$, the multiple correlation coefficient. We may write as the function which is to be made a minimum,

$$f = Nk^2_{1.23\ldots n}.$$

Then we have

$$f = \Sigma(z_1 - \beta_{12.34\ldots n}z_2 - \beta_{13.24\ldots n}z_3 - \cdots - \beta_{1n.23\ldots n-1}z_n)^2. \qquad (8)$$

Taking partial derivatives with respect to $\beta_{12.34\ldots n}$, $\beta_{13.24\ldots n}$, etc., in turn, setting them equal to zero and solving the resulting set of simultaneous equations, yields the values of the β's which will lead to the minimal error of estimate. We may write the answer in convenient form, if total correlation coefficients, r_{12}, r_{13}, etc., have the meaning of Chapter 8, page 122, if we let Δ stand for the determinant,

$$\Delta = \begin{vmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{12} & 1 & r_{23} & \cdots & r_{2n} \\ r_{13} & r_{23} & 1 & \cdots & r_{3n} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ r_{1n} & r_{2n} & r_{3n} & \cdots & 1 \end{vmatrix} \qquad (9)$$

and if we let Δ_{11}, Δ_{12}, etc., stand for the minors obtained by deleting the first row and first column; the first row and second column; etc. With this notation the β's are given by

$$\begin{aligned} \beta_{12.34\ldots n} &= \Delta_{12}/\Delta_{11}, \\ \beta_{13.24\ldots n} &= -\Delta_{13}/\Delta_{11}, \\ \beta_{14.235\ldots n} &= \Delta_{14}/\Delta_{11}, \\ \beta_{1p.23\ldots(p)\ldots n} &= (-1)^p\Delta_{1p}/\Delta_{11}, \\ &\text{etc.} \end{aligned} \qquad (10)$$

The multiple alienation coefficient is (see Pearson, Karl, *Biom.*, v. 8, p. 439, eq. VI, 1912)

$$k_{1.23\ldots n} = \sqrt{\Delta/\Delta_{11}}. \qquad (11)$$

This alienation coefficient may also be obtained from partial alienation coefficients of lower order. Partial alienation coefficients may be defined as constants related to partial correlation coefficients as are multiple and total alienation coefficients related to multiple and total correlation coefficients. Thus, in general, k with any subscript is related to r with the same subscript by the equation,

$$k^2 + r^2 = 1. \qquad (12)$$

Then, as first derived by Yule (*Proc. Roy. Soc.*, A, v. 79, 1907), but using the present notation:

$$k_{1.23\cdots n} = k_{12.34\cdots n}k_{13.45\cdots n}\cdots k_{1n}. \tag{13}$$

The multiple correlation coefficient is,

$$r_{1.23\cdots n} = \sqrt{1 - k^2_{1.23\cdots n}}. \tag{14}$$

This completes the determination of the basic constants of the regression equation as but very simple substitutions are required to get equation (1) having equation (2).

It is sometimes desired to know the "partial" correlation between two variables, that is, the correlation between two variables independent of one or more additional variables. The symbol for the partial correlation of X_1 and X_2 is $r_{12.34\cdots n}$ and may be read, "the correlation between X_1 and X_2 independent of $X_3, X_4, \cdots X_n$," or "the correlation between X_1 and X_2 for constant values of $X_3, X_4, \cdots X_n$." The partial correlation expressed by $r_{12.34\cdots n}$ may be interpreted as an average value of the correlations between those values of X_1 and X_2 which correspond to different assigned values of the remaining $n - 2$ variables. This partial correlation is given by the first or second of the following equalities:[1]

$$r_{12.34\cdots n} = \sqrt{\beta_{12.34\cdots n}\beta_{21.34\cdots n}} = \sqrt{b_{12.34\cdots n}b_{21.34\cdots n}}. \tag{15}$$

The two β's giving the partial correlation coefficient have been called conjugate regression coefficients by Kelley.[2]

Formulas (10) give the value of any β regression coefficient as the quotient of two determinants. A coefficient, β, of a given order may also be evaluated in terms of β-coefficients of lower order, by the following equation:

$$\beta_{12.34\cdots n} = \frac{\beta_{12.45\cdots n} - \beta_{13.45\cdots n}\beta_{32.45\cdots n}}{1 - \beta_{23.45\cdots n}\beta_{32.45\cdots n}}. \tag{16}$$

In equation (16) one (called the unique secondary subscript) and only one of the secondary subscripts appearing in the β in the left-hand member has disappeared from the secondary subscripts in the β's in the right-hand member. Since all but one of the secondary subscripts appear as secondary subscripts in both members the general principle may be illustrated by a β of the second order;

$$\beta_{12.34} = \frac{\beta_{12\ 4} - \beta_{13.4}\beta_{32.4}}{1 - \beta_{23.4}\beta_{32.4}}.$$

[1] See T. L. Kelley, *Statistical Method* (1923); and, G. Udny Yule, *Introduction to the Theory of Statistics* (1912).

[2] T. L. Kelley, *Chart to Facilitate the Calculation of Partial Coefficients of Correlation and Regression Coefficients* (1921).

The first primary subscript in the left-hand member term becomes the first primary subscript in the first and second β's in the numerator of the right-hand member. The second primary subscript in the left-hand member term becomes the second primary subscript in the first and third β's in the numerator. The remaining two primary subscripts of the numerator β's are identical and are the unique secondary subscript. The denominator β's are the third numerator β and its conjugate.

From these general directions it is obvious that there are as many different ways for expressing a regression coefficient of a given order as the order of the coefficient. Thus $\beta_{12.34\ldots n}$ may be expressed in $n - 2$ ways, equation (16) being one such. In practice it is desirable to calculate in at least two ways as a check.

Formula (16) simplifies in case a partial regression coefficient of the first order is being calculated:

$$\beta_{12.3} = \frac{\beta_{12} - \beta_{13}\beta_{32}}{1 - \beta_{23}\beta_{32}} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}. \tag{17}$$

By repeated use of equations (16) and (17) in calculating regression coefficients of a given order from hose of an order one less, every regression coefficient may be obtained. In case determinants are not used equations (16) and (17) will serve as the basis in calculating β's.

For a three- or four-variable problem one of the two methods here given, a method given by Yule,[1] or a method utilizing an alignment chart devised by Kelley[2] will prove serviceable, and for problems involving more than six variables Kelley[3] gives a greatly abbreviated method. The determinantal solution is the most convenient for theoretical work and probably the preferred one for practical work in case of a five- or six-variable problem.

A sample three-variable problem. The steps in the calculations may be illustrated from the following data provided by Mr. V. M. Cady. The data will serve in the problem of the next section involving partial and multiple correlation ratios, as well as in this problem, and thus enable a comparison between correlation coefficient and correlation ratio methods.

The tabulated numbers are the individual observations, thus each tabular number refers to one individual. The values of the first variable

[1] G. Udny Yule, *Introduction to the Theory of Statistics*, chap. XII (1912).

[2] T. L. Kelley, *Chart to Facilitate the Calculation of Partial Coefficients of Correlation and Regression Equations* (1921).

[3] T. L. Kelley, *Statistical Method* (1923, pp. 302–08).

TABLE OF DATA

x_3

	3	4	5	6	7	8	9	10	11
$x_2 = \alpha$	11 8	11 12 11 10 11 9 10 8 10 9 12 6 12 7	13 9 10 9 10 9 11 4 8 7 10 5 10 9 12	7 7 8 9 11 7 7 7 6 8	8 7 8 5	10 9	11		
$x_2 = \beta$		8 7 10 7 7 5 5 2 8	11 6 10 5 10 0 10 3 6 6 6 6 9 3 6 7 7 6 6 7 4 7 7	4 9 12 6 5 5 4	10 7 5	6			9
$x_2 = \gamma$	5	10 4 7 3 7	8 2 11 5 4 7 3 4 0 6 4 2 3 0 3 8	10 4 2 5 5 5 5 6 5 6 6	3 1 5	3 2 0 4 5 6 3	5		

are recorded in the cells of the table. This variable is the rating in "honesty" given by the teachers of certain school children. The second variable is categorical, consisting of three groups. The γ group comprise children rated as the most incorrigible by the school principal or disciplinarian. The β group comprise children rated as of average corrigibility, and the α group the most corrigible children. The third variable is the score in a test measuring the extent to which the children's judgments of the seriousness of offenses correlate with adult judgments of the same items. The distribution of children in the categorical trait would probably be tri-modal, if finely graded measures were available. In the partial and multiple correlation ratio treatment the categorical nature of this variable is no hindrance, but in order to calculate partial and multiple correlation coefficients we must assign numerical values to the categories. We will assign the quite arbitrary values, $\alpha = 3$; $\beta = 2$; $\gamma = 1$. Straightforward calculation yields:

		Variables		
		x_1	x_2	x_3
Cor-	x_1	1.0000	.6076	− .2064
rela-	x_2	.6076	1.0000	− .1715
tions	x_3	− .2064	− .1715	1.0000
Means		6.7573	2.0294	5.4191
σ's		2.9293	.8220	1.3036

$$\Delta = \begin{vmatrix} 1.0000 & .6076 & -.2064 \\ .6076 & 1.0000 & -.1715 \\ -.2064 & -.1715 & 1.0000 \end{vmatrix} = .60182.$$

$$\Delta_{11} = \begin{vmatrix} 1.0000 & -.1715 \\ -.1715 & 1.0000 \end{vmatrix} = .97059.$$

$$\Delta_{12} = \Delta_{21} = \begin{vmatrix} .6076 & -.1715 \\ -.2064 & 1.0000 \end{vmatrix} = .57220.$$

$$\Delta_{13} = \begin{vmatrix} .6076 & 1.0000 \\ -.2064 & -.1715 \end{vmatrix} = .10220.$$

$$\Delta_{22} = \begin{vmatrix} 1.0000 & -.2064 \\ -.2064 & 1.0000 \end{vmatrix} = .95740.$$

$$k_{1.23} = \sqrt{\frac{\Delta}{\Delta_{11}}} = .78744, \qquad r_{1.23} = \sqrt{1 - k^2_{1.23}} = .61639.$$

$$\beta_{12.3} = \Delta_{12}/\Delta_{11} = .58954, \qquad \beta_{13.2} = -\Delta_{13}/\Delta_{11} = -.10530,$$

$$\beta_{21.3} = \Delta_{21}/\Delta_{22} = .59766.$$

$$r_{12.3} = \sqrt{\beta_{12.3}\beta_{21.3}} = .59359.$$

$$\bar{z}_1 = \beta_{12.3}\, z_2 + \beta_{13.2}\, z_3 = .58954\, z_2 - .10530\, z_3.$$

$$\bar{X}_1 = 2.1008\, X_2 - .2366\, X_3 + 3.7761.$$

Thus, for this problem, we find by referring to the z regression equation that the school principal's ratings should be given six times as much weight as the judgment test scores when the two are combined to estimate honesty scores. However, the nominal weights, due to differences in variability, are as 2.1008 is to .2366. The fact that the second coefficient is negative merely means that greater deviation in judgment goes with greater dishonesty, and is thus the sign that one would expect. The correlation between principal's ratings and honesty scores, .6076, is nearly as high as that between honesty scores and a composite of principal's ratings and judgment test scores, since this latter is .6164. This shows that the judgment test has contributed but little that is not already involved in the principal's ratings. Another evidence of this same fact is that the correlation between honesty and principal's ratings independent of the judgment test, .5936, is almost as high as the total correlation between these two measures, .6076. In the next section the relationships without assuming linear regression will be worked out.

GENERALIZED CORRELATION RATIOS

Partial and multiple correlation ratios. From Chapter 8, page 128 we have, in the case of linear regression, the value of

$$\sigma_{1.2} = \sigma_1\sqrt{1 - r_{12}^2}.$$

This is an average of the standard deviations of arrays of X_1's around the linear regression line. Further, we have

$$\sigma_{1.2}^2 = \sigma_1^2 - \sigma_{\bar{x}_1}^2$$

in which $\sigma_{\bar{x}_1}$ is the standard deviation of the means of arrays of X_1's (see page 140) or the standard deviations of the values of X_1 calculated from the regression equation by assigning values to X_2.

From these two equations, we have

$$r_{12} = \frac{\sigma_{\bar{x}_1}}{\sigma_1}.$$

Thus, if regression is linear, the correlation coefficient is the ratio of these two standard deviations. In Chapter 8 this ratio, independent of linearity of regression, is defined as the correlation ratio and is given the symbol η_{12}. Thus,

$$\eta_{12} = \frac{\sigma_{\bar{x}_1}}{\sigma_1},$$

where $\sigma_{\bar{x}_1}$ is still the standard deviation of the means of arrays, but not ordinarily the standard deviation of values of X_1 calculated from a linear regression equation.

Dealing with several variables the same sort of a relation maintains between the multiple correlation coefficient and the multiple correlation ratio. For linear regression we have,

$$\sigma_{1.23\cdots n} = \sigma_1\sqrt{1 - r^2_{1.23\cdots n}}$$

and also,

$$\sigma^2{}_{1.23\cdots n} = \sigma^2{}_1 - \sigma^2_{\bar{x}_1},$$

in which $\sigma_{\bar{x}_1}$ is the standard deviation of the x_1 values estimated by means of the regression equation from a knowledge of the $x_2, x_3, \cdots x_n$ values. Solving for $r_{1.23\cdots n}$, we obtain,

$$r_{1.23\cdots n} = \frac{\sigma_{\bar{x}_1}}{\sigma_1}.$$

Thus, if regression is linear, the multiple correlation coefficient is the ratio of two standard deviations. This ratio maintains significance if the regression is non-linear, but we now give it a new name and symbol. It is called the multiple correlation ratio. Thus,

$$\eta_{1.23\cdots n} = \frac{\sigma_{x_1}}{\sigma_1}. \tag{18}$$

The calculation of $\eta_{1.23\ldots n}$ is straightforward though laborious in case the number of variables and the number of categories per variable is not small. The standard deviation σ_1 is simply the standard deviation of the x_1 measures when included in a single distribution. The magnitude $\bar{x}_1$ is the mean of the x_1 measures lying within a single cell of the multiple correlation surface. Thus, if there are four variables, x_1, x_2, x_3, x_4, and if x_2 has ten classes or categories; x_3, 15; and x_4, 12; there are, all told, 1800($= 10 \times 15 \times 12$) cells, and 1800 different values of $\bar{x}_1$. The standard deviation of these, taking each as many times as there are cases in the cell, is the $\sigma_{\bar{x}_1}$ desired. From the statement just made it is obvious that x_2, x_3, x_4, etc., need not be graded variables. All of the independent variables may be strictly categorical, it only being necessary that the dependent variable x_1 be graded in order to calculate a correlation ratio. It is also obvious that the cell populations, in case of 1800 cells, would be very small, unless a very large total population is dealt with. As a consequence, unless the population is several times the number of cells, there is a very large grouping error, tending systematically to make the multiple correlation ratio too large. Pearson [1] and Student [2] have taken the initial steps in correcting for this error. If

[1] Karl Pearson, "On a Correction to be Made to the Correlation Ratio," *Biometrika*, vol. 8 (1911), p. 254.

[2] Student, "The Correction to be Made to the Correlation Ratio for Grouping," *Biometrika*, vol. 9 (1913), p. 316.

the grouping is too coarse there is an error of a different sort, but, in general, the value of the raw multiple correlation ratio will be more accurate if grouping is so coarse that no cell frequency is less than 2, than if a finer grouping is utilized.

The concept of a partial correlation ratio follows directly from that of a partial correlation coefficient. Formula (13) gives,

$$k_{12.34\,\cdots\,n} = \frac{k_{1.23\,\cdots\,n}}{k_{1n}k_{1(n-1).n}\,\cdots\,k_{13.45\,\cdots\,n}},$$

yielding a partial coefficient of alienation of order $(n - 2)$ in terms of the multiple coefficient of alienation, and partial coefficients of lesser order. A similar relationship holds if non-linear regression is being measured [1] so that we have,

$$\sqrt{1 - \eta^2_{12.34\,\cdots\,n}} = \left\{ \frac{1 - \eta^2_{1\cdot 23\,\cdots\,n}}{(1 - \eta^2_{1n})(1 - \eta^2_{1(n-1)\cdot n})\cdots(1 - \eta^2_{13\cdot 45\,\cdots\,n})} \right\}^{\frac{1}{2}} \quad (19)$$

The same observations as to grouping and as to the categorical nature of the variables hold here as in the case of the multiple correlation ratio. In the case of multiple partial correlation ratios, just as in the case of the total correlation ratios, subscripts before the point are not interchangeable. Thus, in general

$$\eta_{12.3} \neq \eta_{21.3}, \text{ but } \eta_{1.23} = \eta_{1.32}.$$

A sample three-variable problem. We may illustrate formulas (18) and (19) using the data given earlier in this chapter. Let it be required to find the correlation ratio of estimates of honesty upon disciplinarian's records and score in the judgment tests. We have nine classes in x_3 and three classes in x_2, making a total of 27 cells. Of these, seven have zero frequencies and five have frequencies of one each. The grouping in x_3 is not as coarse in the region of the upper and lower ends of the distribution as might be desirable, but, taking it as it stands, we have for the $(x_2 = \alpha,\ x_3 = 3)$ cell a mean $\bar{x}_1 = 9.5$ and a cell frequency of 2; for the $(x_2 = \alpha, x_3 = 4)$ cell a mean $\bar{x}_1 = 9.857$ and a cell frequency of 14; etc., for the remaining cells. The standard deviation of these means, taking each as many times as the population of the cell, is by the usual calculation equal to 1.971 and σ_1 as given on page 145 is 2.929. Thus, $\eta_{1.23} = .6729$. This value may be compared with $r_{1.23}$, which was found on page 145 to equal .6164.

[1] Karl Pearson, "On the Partial Correlation Ratio," *Proc. Roy. Soc.*, A, vol. 91 (1915).

Let it be required to find the partial correlation ratio of x_1 upon x_3 when x_2 is constant. Formula (19) states,

$$1 - \eta^2_{13.2} = \frac{1 - \eta^2_{1.23}}{1 - \eta^2_{12}}.$$

Since η_{12} by the calculation of Chapter 8 is found to equal .5900, we obtain $\eta_{13.2} = .4009$. That is to say that, independent of the bearing or relationship of disciplinarian's records, there is a correlation ratio of .4 between teachers' estimates of honesty and test score. This value is probably somewhat too large on account of the grouping error.

When a student first gains a knowledge of the general principles of partial and multiple correlation and added power in analysis resulting from their use, there is a danger of misuse. As a general principle, it is better to separate experimentally, if possible, the various factors than to determine their separate relationships statistically by means of partial correlation. In many cases the experimental and statistical analyses can both be used, thus supplementing each other.

The most subtle danger is that the tyro will interpret partial coefficients of correlation as measuring causal relationships, whereas in fact the cause of a relationship is no more betrayed by a partial coefficient in the case of a number of variables than by a total coefficient of correlation in the case of two.

There is less danger in interpreting a multiple coefficient of correlation as it is readily placed in the same category as the total correlation coefficient. The care to be exercised here should be to see that much of significance is not lost by treating regression relationships as linear. This point may be numerically tested by seeing if the multiple correlation coefficient is nearly as large as the multiple correlation ratio. If several variables are involved, the process would be laborious and the probable error of the difference

$$\eta^2_{1.23 \cdots n} - r^2_{1.23 \cdots n}$$

would undoubtedly be large, so that one is thrown back upon the necessity of making supplementary investigations and exercising his good judgment as to whether linear relationships may reasonably be assumed.

If relationships are linear, if the accuracy consequent to the size of the population dealt with is appreciated, if causal relationships are not assumed, and if conclusions are not extended to populations which are not homogeneous with the sample, then there is full warrant for use of the powerful analytical devices of multiple and partial correlation.

CHAPTER X

CORRELATION OF TIME SERIES

By WARREN M. PERSONS

Each item of a time series of statistics is an aggregate or average or relative number applying to a definite interval or point of time. Unless otherwise specified, the several items of a series are understood (1) to refer to equal time units, (2) to be consecutive in time, and (3) to be constructed according to a fixed criterion or standard. Illustrations of time series are: the population of continental United States at each census, the aggregate pig-iron production per month of a representative number of blast furnaces, average rates on commercial paper each week or month, index numbers of prices on the first of each month.

The items of time series must be defined for selected units of time — the week, the month, the quarter, the year. Because of this fact the items are *ordered* in time and therefore are affected by the same or related influences during adjacent time-intervals. In other words, each time series is a function of time and, presented graphically, has a characteristic conformation. In this respect time series differ from other series of statistics, such as the wage rates of different individuals or the populations of different countries at a given time.

Four types of variations are commonly found in the ordered items of time series. They are, first, variations which occur within each year as a consequence of the round of the seasons by which the items for certain weeks or months are regularly higher or lower than those for other weeks or months of the year; second, a long-time movement or secular trend covering a considerable period of years by which the average size of the items makes a permanent gain or suffers a permanent loss; third, irregular fluctuations resulting from wars, panics, strikes, etc.; and fourth, wave-like or cyclical movements — which may or may not be periodic — connected with the ebb and flow of business.

If our problem is to ascertain the relationship between two series ordered in time it is of little avail (or actually misleading) to compute the coefficient of correlation from pairs of the actual items. In case the two series possess definite trends or seasonal variation the coefficient of correlation for the items will yield a value different from zero. Having found such a coefficient we would be unable to say what contributed most largely to the result — similar (or diverse) trends, seasonal variations, cyclical movements, or irregular fluctuations. Generally in the comparison of two time series it is the relation between their respective cyclical variations that is most interesting from the practical point of view. In order that this relation may be set forth either graphically or by use of the correlation coefficient, it is necessary to remove from the actual items that portion of their values ascribable to secular trend and seasonal variation. It would be desirable also to eliminate the irregular fluctuations, but this appears to be impossible in general because, by definition, such fluctuations are unsystematic. Our problem, then, resolves itself into two parts, first, the measurement and elimination of seasonal variation and secular trend from each series under investigation, and, second, the measurement of the correlation between the two series of items thus " corrected."

SEASONAL VARIATION

The problem of measuring seasonal variation is that of isolating from a time series a *typical* movement having a period of one year. In attacking this problem it is essential to adapt our methods to the material which we are using: time series in which seasonal variations, unlike those of the physical sciences, do not usually occur with a high degree of uniformity. Furthermore, the problem is greatly complicated, especially when economic series are our material, both by numerous irregular fluctuations and by lack of homogeneity over a long interval of time.

Suppose that there is given a series of economic statistics, the monthly items of which we shall designate by y, and that we seek to measure the seasonal variation. The procedure is as follows: [1]

First, for every month, except the first, calculate the *link relative*, which is the ratio of each item of the series to the preceding item, or $\frac{y_i}{y_{i-1}}$. In this way each January item is expressed as a percentage of the preceding December, each February item as a percentage of the preceding January, and so on.

Second, arrange the January link relatives in a frequency table, the February link relatives in another adjacent frequency table, and so

[1] The method is, essentially, the one described in the *Review of Economic Statistics*, January, 1919, pp. 18–31, of the article by W. M. Persons on "Indices of Business Conditions."

on for the remainder of the months. An illustration of the several frequency tables thus secured for the monthly rate on 60–90 day commercial paper in New York, January, 1890–January, 1917, is given in Figure 9.

Third, find the *median* of each frequency table.[1]

Fourth, express each median link relative as a percentage based on January by progressively multiplying or "chaining" the medians. Thus, January is taken as 100; the product of this 100 by the February median (percentage) gives the February item of the chain series; the product of this result by the March median gives the March item; and in a similar way we get successively all of the items to December of the chain series. If the computed December item is multiplied by the January median, we shall find that the result will not ordinarily be 100 but some other percentage. That is, the process of chaining medians, or any averages except the geometric means of a series in which the initial and final items are identical,[2] gives rise to some discrepancy.

Fifth, distribute the discrepancy among all the items of the chain series, using either a geometric or an arithmetic basis for adjustment. The object of the adjustment is to make the computed January chain relative equal 100.

Finally, alter proportionally the revised relatives thus secured so that their arithmetic mean shall be 100. These final figures are the *adjusted monthly indexes of seasonal variation.*

Expressed in mathematical symbols the process of finding the adjusted indexes of seasonal variation is as follows:

Let $r_1, r_2, r_3, \cdots r_{12}$ be the medians of the link relatives (expressed as decimals).

Let $100, c_2, c_3, \cdots c_{12}$ be the chain relatives obtained by progressive multiplication, with January as 100.

Then, $c_2 = 100\, r_2,\ c_3 = c_2r_3,\ c_4 = c_3r_4, \cdots c_{12} = c_{11}r_{12}$.

In this series of equations r_1 and c_1 do not appear. If we *compute* a value for January (represented by c_1) we shall have $c_1 = c_{12}r_1$. The

[1] The average of three or four central items might appropriately be taken instead of the median. In case of a very large number of items in the frequency tables having a clearly marked class of concentration, the measurement corresponding to that class, the mode, might appropriately be taken as the typical seasonal relative. Economic series are not, however, sufficiently long or homogeneous to make the use of the mode practicable.

[2] The discrepancy for a chain series of geometric means in which y_o and y_f are the initial and final actual items (referring to the same calendar month) of the series is $\sqrt[n]{\frac{y_f}{y_o}}$, where n is the number of years covered by our monthly series.

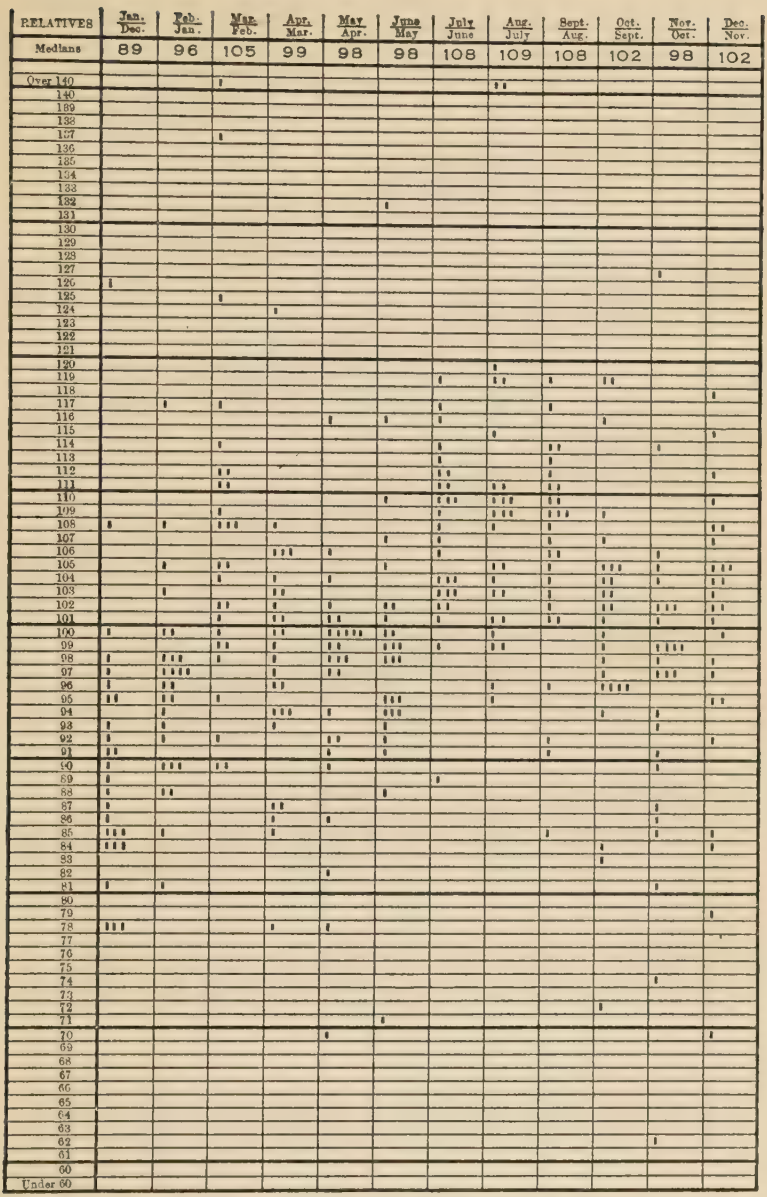


FIG. 9. FREQUENCY DISTRIBUTIONS, ARRANGED BY MONTHS, OF RELATIVES FOUND BY EXPRESSING THE RATE ON 60–90 DAY COMMERCIAL PAPER IN NEW YORK EACH MONTH AS A PERCENTAGE OF THE ITEM FOR THE PRECEDING MONTH: JANUARY, 1890–JANUARY, 1917

product $c_{12}r_1$ will not, in general, give 100 but some discrepancy in excess or defect of 100.

Assuming that the discrepancy may be appropriately distributed by applying a constant factor $1 + d$ to the monthly medians of link relatives, we have

$$100\,(1 + d)^{12} = c_{12}r_1,$$

from which $1 + d$ may be computed.

If, now, we adjust the chain relatives as follows:

$$100, \frac{c_2}{1 + d}, \frac{c_3}{(1 + d)^2}, \cdots \frac{c_{12}}{(1 + d)^{11}}, \frac{c_{12}r_1}{(1 + d)^{12}},$$

the last term will be 100 and the discrepancy thus disappears.

The computation is greatly facilitated by the use of logarithms. A scheme for the computation is given in Table I.

TABLE I. COMPUTATION OF INDEXES OF SEASONAL VARIATION

(Rates on 60–90 Day Paper, Jan., 1890–Jan., 1917) [1]

Median of Link Relatives	Logarithm of Median Link Relative (Log r)	Months	Logarithm of Chain Index with Jan. as Base (Log c)	Logarithm of Adjustment Factor[2] Log $(1+d)t$	Logarithm of Adjusted Chain Index with Jan. as Base (Log a)	Adjusted Chain Index with Jan. as Base (a)	Index of Seasonal Variation (Average for Year = 100) (s)
89	1.9494	Jan.	2.0000	0.0000	2.0000	100.0	97.0
96	1.9822	Feb.	1.9822	0.0036	1.9786	95.2	92.4
105	2.0212	Mar.	2.0034	0.0072	1.9962	99.1	96.1
99	1.9956	Apr.	1.9990	0.0109	1.9881	97.3	94.4
98	1.9912	May	1.9902	0.0140	1.9762	94.7	91.9
98	1.9912	June	1.9814	0.0181	1.9633	91.9	89.2
108	2.0334	July	2.0148	0.0217	1.9931	98.4	95.4
109	2.0374	Aug.	2.0522	0.0253	2.0269	106.4	103.3
108	2.0334	Sept.	2.0856	0.0290	2.0566	113.9	110.5
102	2.0086	Oct.	2.0942	0.0326	2.0616	115.2	111.8
98	1.9912	Nov.	2.0854	0.0362	2.0492	112.0	108.6
102	2.0086	Dec.	2.0940	0.0398	2.0542	113.3	110.0
89	1.9494	Jan.	2.0434	0.0434	2.0000	.....	.....
Arithmetic average	......	....	......	......	......	103.1	100.0

[1] Based on data published in the *Review of Economic Statistics*, January, 1923, p. 28.

[2] Time (t) measured in months from January. Log $(1 + d) = \frac{1}{12}$ (0.0434). To adjust by an arithmetic process we would of course apply the equation of condition:

$$d' = \frac{1}{12}\,(c_{12}r_1 - 100).$$

There is theoretical preference for geometric distribution of the discrepancy, but the two methods are unlikely to yield significantly different results in ordinary practical cases.

The advantages of using the method just outlined, compared, for instance, with the method of arithmetic means of the actual January items, February items, etc., are as follows:

1. The frequency distributions of link relatives enable one to judge the degree or regularity of month-to-month (or seasonal) changes. The closer the grouping of relatives for a given month about any value, the more pronounced and significant is the seasonal movement for that month.

2. The use of the median (or average of central items) is a device by which the influence of extremely large non-seasonal variations (such, for instance, as the sudden decline of money rates after the defeat of Bryan in November, 1896, or the rapid rise in August, 1914) may be greatly moderated.[1]

3. It is possible to utilize non-homogeneous statistical series in the measurement of seasonal variation. For instance, link relatives for the bank clearings of 50 representative cities for one interval may be combined with link relatives for the clearings of 100 representative cities for another interval. Likewise, the method may be used when only adjacent *pairs* of items are comparable.

Series of economic statistics sometimes occur, covering an interval of 10 or 15 years, in which the items are strictly homogeneous and not affected by large irregular fluctuations. In such cases the method of arithmetic or geometric means of all the actual items for corresponding months may be used in determining indexes of seasonal variation.[2]

If we assume (as does W. L. Hart in "The Method of Monthly Means for Determination of a Seasonal Variation," *Jour. Amer. Statis. Ass'n*, September, 1922, pp. 341–349) that monthly series of economic data are described with sufficient accuracy by trigonometric functions, such as $f(t) = A + B \sin t\,(30°) + C \sin t\left(\frac{30°}{n}\right)$, then it may be demonstrated that the method of monthly means gives the correct seasonal variation for such functions. Thus, in the function given above, where

t = time in months,
n = an integer (two or greater),
$B \sin t\,(30°)$ = seasonal variation component, period of 1 year,
$C \sin t\left(\frac{30°}{n}\right)$ = cyclical variation component, period of n years, the arithmetic mean of the values of the function $f(t)$ for corresponding months (each July, for

[1] W. L. Crum, "The Use of the Median in Determining Seasonal Variation," *Jour. Amer. Statis. Ass'n*, March, 1923.

[2] If in the median method (previously described) geometric averages of the link relatives had been used (instead of medians) and these averages had been multiplied together progressively to secure a continuous series of relatives with a fixed base, the result would have been identically the same as that secured by expressing the original items in terms of a fixed base and then taking the geometric averages of these percentages.

instance) during a complete cycle (n years) will be $A + B \sin t\,(30°)$. Since the arithmetic mean of all values of the series is A, we get the seasonal variation components by subtracting the average for all months from the average of each of the 12 groups of corresponding months.

Instead of building up an additive function suppose we construct a product function as follows:

$$f(t) = A \cdot B^{\sin t\,(30°)} \cdot C^{\sin t\left(\frac{30°}{n}\right)}.$$

For this function the *geometric* mean of the values for corresponding months during a complete cycle will give the correct seasonal variation. The demonstration of the statement just made follows immediately from the theorem cited by Mr. Hart.[1]

Comparison of the results obtained by the median and the arithmetic average methods of securing indexes of seasonal variation for an illustrative case — the longest monthly series available, rates on commercial paper since 1866 — is given in Figure 10.[2] Examination of the actual data, 1866–1913, showed that there were the following four periods with characteristic fluctuations:

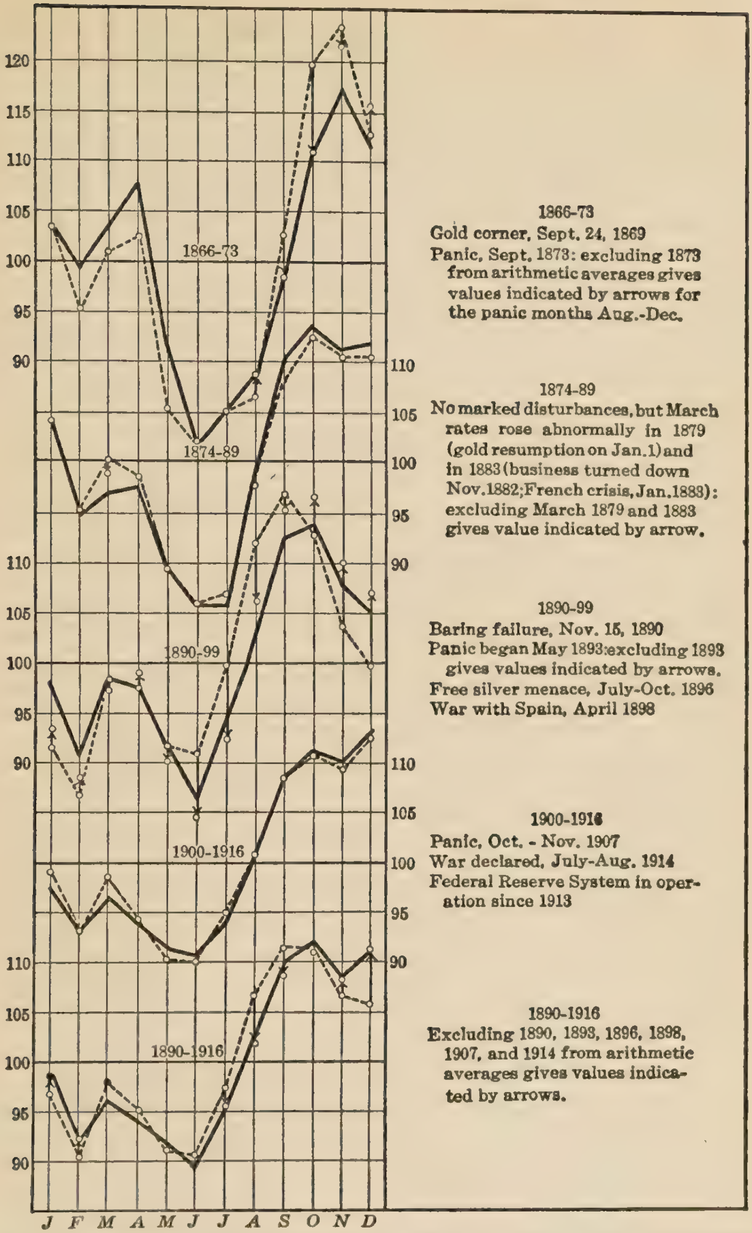
1866–73:	Violent fluctuations on a high level
1874–89:	Narrow fluctuations on a lower level
1890–99:	Marked irregular fluctuations
1900–13:	Very few irregularities

Two sets of seasonal indexes, one based on medians of the link relatives, the other on arithmetic averages of the actual items, were computed for each of these four periods and, in addition, for the period 1890–1916. The five pairs of indexes are presented graphically. It is evident that the two sets of indexes for the periods 1874–89 and 1900–13, in which there are no important irregular variations, almost coincide; while the pairs of indexes for the disturbed periods 1866–73 and 1890–99 differ widely. However, if the erratic and clearly non-seasonal fluctuations of the panic months of the autumn of 1873 be omitted from the arithmetic averages, the resulting indexes for 1866–73 will be very much closer to the indexes based upon medians.[3] Likewise, if the panic year 1893 with its wide fluctuations be omitted from the arithmetic averages for 1890–99, the resulting indexes will closely approximate the median indexes.

[1] Bôcher, *Annals of Mathematics*, Second Series, vol. 7 (1906), p. 135, Formula 63.

[2] The data upon which these computations are based may be found in the *Rev. of Economic Statis.*, Jan., 1923, p. 28. The table has been revised slightly since the computation of the seasonal indexes here quoted.

[3] For instance, exclusion of 1873 from the arithmetic averages results in altering the seasonal index for October from 120.4 to 111.7; the index based upon the medians for all years is 110.9



1866-73
Gold corner, Sept. 24, 1869
Panic, Sept. 1873: excluding 1873 from arithmetic averages gives values indicated by arrows for the panic months Aug.-Dec.

1874-89
No marked disturbances, but March rates rose abnormally in 1879 (gold resumption on Jan. 1) and in 1883 (business turned down Nov. 1882; French crisis, Jan. 1883): excluding March 1879 and 1883 gives value indicated by arrow.

1890-99
Baring failure, Nov. 15, 1890
Panic began May 1893: excluding 1893 gives values indicated by arrows.
Free silver menace, July-Oct. 1896
War with Spain, April 1898

1900-1916
Panic, Oct. - Nov. 1907
War declared, July-Aug. 1914
Federal Reserve System in operation since 1913

1890-1916
Excluding 1890, 1893, 1896, 1898, 1907, and 1914 from arithmetic averages gives values indicated by arrows.

FIG. 10. COMPARISON OF THE INDEXES OF SEASONAL VARIATION FOR RATES ON 60–90 DAY COMMERCIAL PAPER COMPUTED FOR VARIOUS PERIODS BY (A) THE MEDIAN METHOD AND (B) THE ARITHMETIC AVERAGE METHOD

|—|—|—|— Median Method - - - - Arithmetic Average Method

Finally, the selective effect of using the median method is clearly shown when the period 1890–1916, having both highly disturbed and comparatively undisturbed sections, is taken. Omission of the disturbed years 1890, 1893, 1896, 1898, 1907, and 1914 (with their non-seasonal fluctuations) from the arithmetic averages gives results which agree (within the limits of accuracy of our data) with those obtained by the median method in which all of the data were utilized. The seasonal indexes for the period 1890–1916, given in Table II, are based upon (A) the arithmetic averages for 1890–1916, (B) the arithmetic averages for 1890–1916 excluding 6 disturbed years, and (C) the median link relative method for 1890–1916. The absolute sum of the differences between corresponding indexes A and B is 20.9, and that between indexes B and C is only 9.7. It is clear, therefore, that the non-seasonal movements of disturbed years distort the seasonal indexes obtained by the method of arithmetic averages. Moreover, it may be pointed out, the extremely high or low items of an economic series are precisely the ones concerning which the accuracy of the data is in greatest doubt.

TABLE II. INDEXES OF SEASONAL VARIATION[1]

METHOD		JAN.	FEB.	MAR.	APR.	MAY	JUNE	JULY	AUG.	SEPT.	OCT.	NOV.	DEC.
Arithmetic average	A[2]	96.6	90.7	97.4	95.0	91.2	90.7	97.2	106.3	111.7	111.5	106.5	105.6
	B[2]	98.2	92.4	97.7	95.0	91.1	90.2	95.4	101.8	108.5	111.0	107.8	111.0
Median . . .	C[3]	98.8	92.1	95.9	94.1	91.5	88.9	95.2	102.9	110.2	111.5	108.4	110.7

[1] Computed from data published in *Indices of General Business Conditions*, by W. M. Persons, pp. 98–99.

[2] Indexes "A" are based upon the 27 years 1890–1916; indexes "B" are based upon the same period excluding the disturbed years 1890, 1893, 1896, 1898, 1907, 1914.

[3] Indexes "C" differ somewhat from those given in the last column of Table II because the latter are based upon the revised data published in the *Rev. of Economic Statis.*, Jan., 1923, p. 28.

SECULAR TREND[1]

The problem of measurement of the secular trend of a time series is that of fitting a straight line or curve to the graph obtained by plotting times as abscissas and the items as ordinates. In computing the trend for a monthly series extending over a number of years it is not necessary to work with monthly items; a series of annual averages may be substituted and much unnecessary computation avoided. Usually we may assume that the trend is a straight line; in other words, that if the trend were the only source of variation in the series, the actual change in value between two consecutive months would be strictly constant.

[1] For a discussion of the general problem of curve fitting see Chapter IV.

In such a case the particular straight line which represents the trend is the one which best "fits" the data, the criterion of fit being that the sum of the squares of the deviations[1] of the points corresponding to the data from the line shall be a minimum.

To determine this straight line for a series of annual items y_1, y_2, y_3, $\ldots y_n$, proceed as follows: Measure time (t) from the central item of the series, or, if the number of items is even, from the point midway between the pair of central items. For an odd number of annual items the abscissas would be $\cdots, -3, -2, -1, 0, +1, +2, +3, \cdots$, and the unit one year; for an even number of items the origin would be midway between the central pair and, to avoid fractions in the computation, the abscissas would be $\cdots, -5, -3, -1, +1, +3, +5, \cdots$, and the unit one-half year. If the origin for time be so chosen, the annual increment of the line of trend will be $\frac{\Sigma yt}{\Sigma t^2}$; and the y-intercept will be $\frac{\Sigma y}{n}$, or the arithmetic mean of the items.[2] The slope of the line of secular trend can easily be found, therefore, by the following computations: First, multiply the original items by a series of integers denoting units of time, half of which are positive in sign and half negative; second, find the algebraic sum of the products thus secured; third, find the sum of the squares of the series of integers; and, fourth, take the ratio of the first of these sums to the second.

The object of finding a line or curve of secular trend is primarily to secure a function describing the general movement of a time series for a completed historical interval, covering, if possible, several business cycles. The function which best fits the data of a completed historical interval ending with the present, however, is not necessarily the best function for estimating the *future trend.* For estimating future trends

[1] The deviations mentioned are measured parallel to the axis of ordinates.

[2] y = annual average of monthly items.

t = time in years measured from the center of the period covered.

n = number of years in period.

$\frac{\Sigma y}{n}$ = ordinate of trend for center of period.

$\frac{1}{12} \cdot \frac{\Sigma yt}{\Sigma t^2}$ = monthly increment of trend.

$\frac{\Sigma y}{n} \mp \frac{12n-1}{2} \cdot \frac{\Sigma yt}{12\,\Sigma t^2}$ = the ordinates of secular trend for the first and last months respectively of the period.

When the number of years is even, substitute $\frac{t}{2}$ for t in the above formulas.

both the type of function and the interval should be selected with reference to that specific problem.

ELIMINATION OF SEASONAL VARIATION AND SECULAR TREND

Assuming that we have measured satisfactorily both seasonal variation and secular trend, the next problem is to " correct " for these factors the original items of the time series. Let the original items be represented by y, corresponding indexes of seasonal variation by s, and corresponding ordinates of secular trend (in terms of the original unit of tons, dollars, etc.) by o. If there were no irregular or cyclical variations affecting the series, each item would be represented by so, which we may appropriately designate the "normal" value.[1] Since the original item is represented by y, the expression $y - so$ is the deviation from "estimated normal," and $\frac{y - so}{so}$ or $\frac{y}{so} - 1$ is the relative deviation from estimated normal. This is the formula according to which we compute our " corrected " series.[2]

If it be desired to correct the original items for seasonal variation alone, the form $\frac{y}{s}$ may be used, the result being expressed in terms of the units of the original series.

COMPARISON AND CORRELATION OF CORRECTED SERIES

It is now possible to compare the cyclical variations of two or more " corrected " series, and the chief means available is the graphic method. If we are interested in comparing the relative violence of fluctuations of the various series, the corrected items (as described above) of each series should be plotted on translucent paper with time measured in months as abscissas. By superimposing one chart on another (their horizontal axes coinciding) over a glass plate illuminated from beneath, the similarities or differences between the two can be estimated. A comparison of these corrected series is given in Figure 11.

In case we are not interested in the relative amplitudes of fluctua-

[1] This is based on the assumption that seasonal variation increases proportionally with the ordinate of secular trend.

[2] In case we conceive the "normal" to be the ordinate of secular trend, the formula becomes $\frac{y - so}{o}$. In actual practice, where the seasonal indexes range within 10 or 15 per cent of the average, the results obtained by the two formulas would not be significantly different.

tions, but merely in the timing or *lag* of one series with respect to another, or the relative *shapes* of the curves, the various corrected series should be expressed in units selected with the object of making the cyclical fluctuations of like magnitude. An appropriate unit for this purpose is the standard deviation of the corrected items. Therefore, before comparing the curves for lag or relative shape, the items of each corrected series are divided by the standard deviation of that series. The items thus expressed are called *cycles*. A comparison of three series of cycles is given in Figure 12.

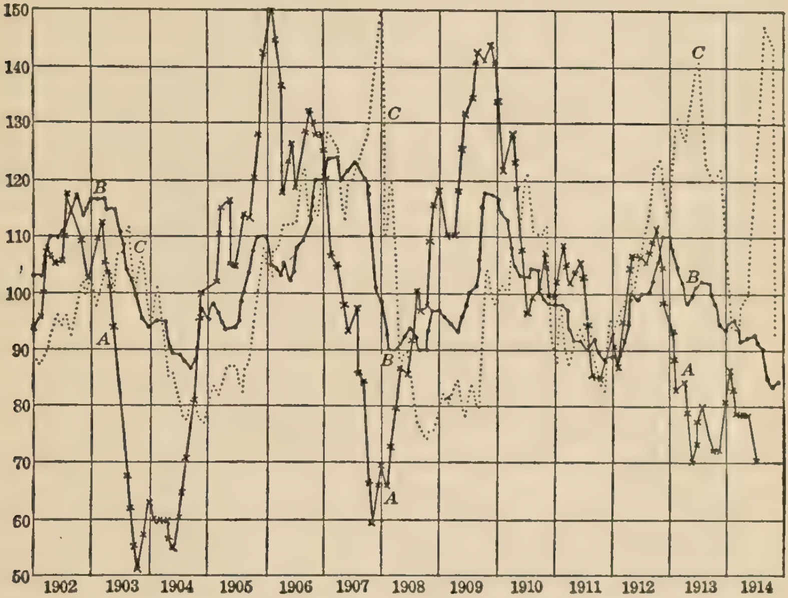


FIG. 11. COMPARISON OF THREE SERIES OF ECONOMIC STATISTICS CORRECTED FOR SEASONAL INFLUENCES AND LONG-TIME TREND

A. Index of the Prices of Twenty Industrial Common Stocks.
B. Ten Commodity Price Index.
C. Rates on Commercial Paper.

(Expressed as percentages of the ordinates of linear secular trend.)

The amount of lag for the highest degree of correlation between two series, or best " fit " when one curve is superimposed upon another, can be determined only approximately by inspection. In order to get numerical measures of the relative goodness of fit for various lags we must resort to the coefficient of correlation (see Chapter VIII). The process is as follows: First, compute the coefficient of correlation for the pairs of items which, according to inspection of the charts, appear to be most highly correlated; and, second, compute other coefficients for lags of both greater and less amount. When some pairing gives a higher coefficient than adjacent pairings, the degree of lag for maximum correlation is indicated.[1] Thus, the coefficients of correlation between the cycles of pig iron production and interest rates, monthly, 1903–14, with a lag of interest rates of 0, 3, 4, 5, 6, 7, 8, 9, and 12 months are, respectively: + .34, + .67, + .72, + .75, + .75, + .73, + .70, + .65, and + .45. These coefficients indicate that the maximum correlation is for a lag of 5 or 6 months in interest rates, a fact which is shown graphically by plotting the coefficients as ordinates and time as abscissas (Figure 13).

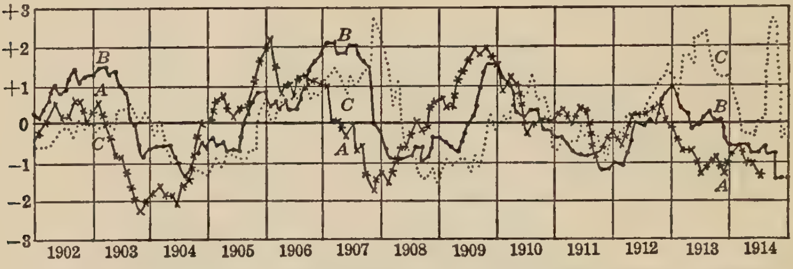


FIG. 12. COMPARISON OF THREE SERIES OF ECONOMIC STATISTICS CORRECTED FOR SEASONAL INFLUENCES AND LONG-TIME TREND

A. Index of the Prices of Twenty Industrial Common Stocks.
B. Ten Commodity Price Index.
C. Rates on Commercial Paper.

(The percentage deviations of Figure 11 are expressed in terms of their respective standard deviations as units.)

The "probable errors" of coefficients of correlation (or other constants) computed from time series of economic statistics do not have the usual meaning. The theory of probability does not apply to our data because,

[1] If in each of two series (x and y) we take deviations from the straight line of secular trend, the sums of such deviations will be zero, or $\Sigma(x - o_x) = o$, and $\Sigma(y - o_y) = o$. If in addition these deviations are measured in terms of their respective standard deviations and we denote the resulting ratios by x' and y', the coefficient of correlation becomes $\frac{\Sigma x'y'}{n}$. In other words, the coefficient of correlation is the arithmetic mean of products of corresponding deviations (when measured in terms of the standard deviation) from the line of secular trend.

For percentage deviations (instead of actual deviations) from the line of secular trend, corrected for seasonal variation, and expressed in terms of their standard deviations (denoted by x'' and y''), the coefficient of correlation is an extremely close approximation to $\frac{1}{n}\Sigma x''y''$.

first, any past period that we select for study is, in fact, a special period with characteristics distinguishing it from other periods, and so cannot be considered a "random" selection; second, the individual items of the series are not chosen independently, but they constitute a *group* of successive items with a characteristic conformation. Consequently, the "probable error" of 0.03 in the coefficient of correlation quoted above does not indicate, as one would conclude from the theory of probability, that if we compute a coefficient from "any" other actual period, the chances are equal that it will be between + 0.72 and + 0.78. In fact, the significance of the probable error of a constant computed from time series is not known.

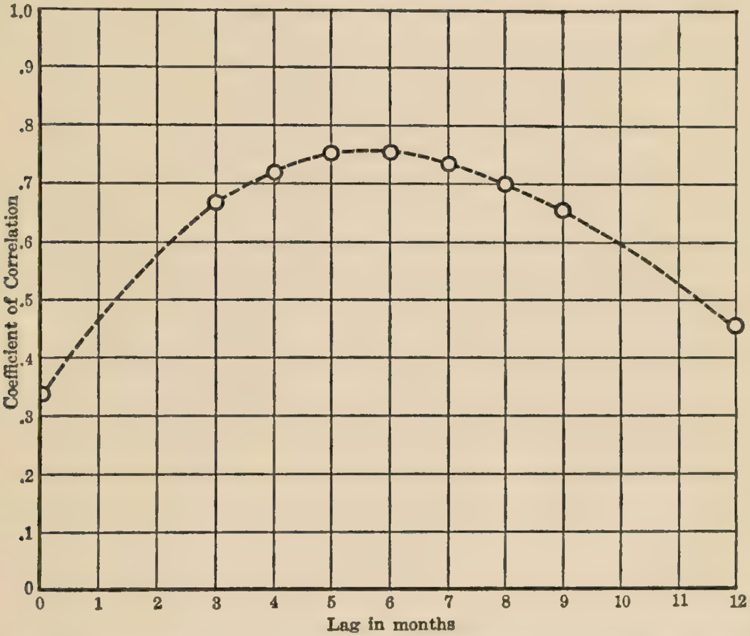


FIG. 13. COEFFICIENTS OF CORRELATION BETWEEN "CYCLES" OF PIG IRON PRODUCTION AND "CYCLES" OF INTEREST RATES ON 60-90 DAY COMMERCIAL PAPER, FOR VARIOUS DEGREES OF LAG IN INTEREST RATES

Coefficients of correlation, although useful in determining lag of time series, are after all merely averages. The specific relationships between two time series are much more adequately set forth by charts than by numerical measures. Also, there is great danger that coefficients based on time series may be wrongly interpreted. For instance, a high coefficient may result if two series fit their secular trends badly and the badness of fit in the two cases is similar.[1] The precise nature of the correlation is often evident from charts when it is not revealed by coefficients.

It is possible, if we are willing to dispense with the graphical comparison of the cycle charts, to eliminate the secular trends of two series and find the coefficient of correlation all in one step by the use of the methods of partial correlation.

Suppose, in fact, that the seasonal variations have already been eliminated and that the resulting series are x' and y'. We consider the problem as one involving three variables, x', y', and t; and the desired correlation between the x' and y', corrected for secular trend, will be equal to

$$r_{x'y'\cdot t} = \frac{r_{x'y'} - r_{x't}.r_{y't}}{\sqrt{1 - r^2_{x't}}\sqrt{1 - r^2_{y't}}},$$

which is the partial correlation coefficient of x' on y' independent of t. Moreover, this value is not altered by the existence or amount of lag: it is at once the desired maximum correlation. It should be remembered, however, that this coefficient deals with deviations rather than the percentage deviations involved in the cycles; but it can be shown that this fact has little bearing on the results.[2]

Coefficients of correlation between first differences of series of corrected monthly items or annual figures are valuable when we desire to measure the similarity or dissimilarity of month-to-month or year-to-year changes, for instance, in investigating the year-to-year movements of the prices and production of crops. The correlation of the second and higher differences, the "variate difference method,"[3] has been proposed as a means of ascertaining the correlation between time series, but the method is based upon assumptions which cannot be retained

[1] Such is the result for most economic series covering both the period of declining prices previous to 1897 and the period of rising prices following that year. Nearly all economic series dip below the linear trend in the nineties so that a correlation coefficient between their deviations would indicate that fact rather than the general correspondence of their fluctuations. (See W. M. Persons, "Construction of a Business Barometer," *Amer. Economic Rev.* (Dec., 1916), p. 755, and H. L. Moore, *Economic Cycles: their Law and Cause*, p. 123.)

[2] W. L. Crum, "A Special Application of Partial Correlation," *Quar. Pub. Am. Stat. Assoc.* (December, 1921), pp. 949–52.

[3] See W. M. Persons, "On the Variate Difference Correlation Method and Curve-Fitting," *Quar. Pub. Amer. Statis. Assoc.*, June, 1917; and G. U. Yule, "The Problem of Time Correlation, with especial reference to the Variate Difference Correlation Method," *Jour. Roy. Statis. Soc.*, July, 1921. In order to be sound the method would have to be altered to allow for the correlation between the original items of a series. Since this was written an article by Karl Pearson and E. M. Elderton "On the Variate Difference Method" has appeared in *Biometrika* for March, 1923, in which the defective character of the variate difference method, as originally proposed, is admitted.

even in the simplest problems, and analyses have shown it to be of little or no value for our purpose.

We may summarize our conclusions as follows: The most satisfactory method of setting forth the relationships between time series of economic statistics is, first, to compute and compare their respective indexes of seasonal variation; second, to compute and compare their lines or curves of secular trend; and, third, to correct the original items for seasonal variation and secular trend, and compare the resulting graphs. Coefficients of correlation for the various possible pairs of items of time series are useful mainly as a basis for judging the lag of one series with respect to another.

CHAPTER XI

PERIODOGRAM ANALYSIS[1]

By W. L. CRUM

HARMONIC ANALYSIS FOR KNOWN PERIODS

The nature of periodicity. The study of a statistical series ordered in time includes, in many cases, an investigation of the existence and extent of periodic variation. In many problems of natural science the periodic portion of the fluctuation is fairly prominent, and the examination of periodicity is direct and simple; but in the problems of social science, and indeed in many of the more complex of the problems of natural science, the determination of periodicity is considerably more difficult. For these less obvious cases the methods of Fourier Analysis,[2] supplemented by Schuster's Periodogram,[3] are coming into wide use. We present in this chapter a summary description of these methods, and certain critical remarks on their use and applicability.

We consider first, to bring out the basic properties of a periodic function, the elementary periodic expression

$$y = a_0 + a \sin\left(\frac{2\pi}{p} t + \alpha\right), \tag{1}$$

in which p is the period, α is the phase, a is the amplitude, and a_0 is the mean value in any period. It is clear that y has its mean value, a_0, when

$$t = -\frac{p}{2\pi}\alpha + k\pi,$$

where k is an integer; and that a is the maximum deviation of y from a_0, and occurs when

$$t = -\frac{p}{2\pi}\alpha + \left(2k - 1\right)\frac{\pi}{2}.$$

[1] The writer of this chapter is indebted to Professor Allyn A. Young for many valuable suggestions.

[2] W. E. Byerly, *Fourier's Series* (Ginn, 1893), chap. II–IV.

[3] D. Brunt, *The Combination of Observations* (Cambridge University Press, 1917), chap. XII.

The graph of the curve is shown in Figure 14. The well-known geometrical construction of y as the projection, on a fixed line OY, of a point A_y revolving with constant angular velocity about a fixed point A_0, is represented in Figure 15. These two figures exhibit the essential features of the simple periodic function upon which our analysis is based.

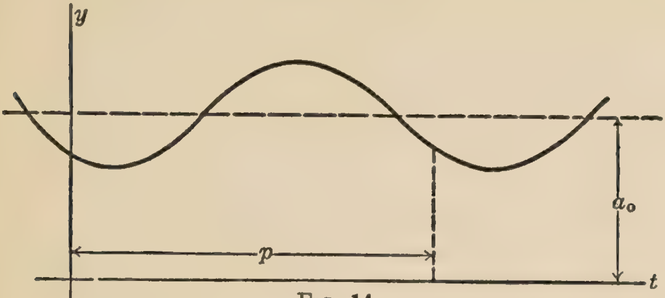


FIG. 14

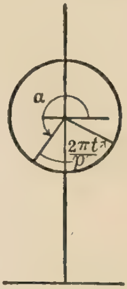


FIG. 15

A familiar trigonometric transformation converts (1) into

$$y = a_0 + a_1 \cos \frac{2\pi}{p} t + b_1 \sin \frac{2\pi}{p} t, \tag{2}$$

where

$$a_1 = a \sin \alpha,\ b_1 = a \cos \alpha, \tag{2 a}$$

from which

$$a = \sqrt{a_1^2 + b_1^2},\ \alpha = \tan^{-1} \frac{a_1}{b_1}. \tag{3}$$

For problems in which the period is known in advance, the method consists in determining the coefficients of (2), and then calculating the amplitude and phase with the help of (3). The four elements, p, a_0, a, α, are then known, and the periodic term is completely determined.

The analysis for a single known period. We suppose that we are concerned with a simple physical problem possessing a known period, p. If there are given three pairs of values of t and y, provided only that the values of t are all different, substitution will give three algebraic equations from which to get a_0, a_1, b_1. Then we can calculate the phase and amplitude by use of (3), and the problem is completely solved.

In general, however, we should have data for more than three pairs of values; and it is then impossible to make the function fit all the observations exactly. In such a case, the coefficients, a_0, a_1, b_1, are determined by the method of least squares, in order that the function may fit all the data as nearly as possible.

Let it be assumed that there are n observations covering a single period, and that these observations are equally spaced in time at intervals of p/n. Such a series is shown in Table I. Substitution in (2) gives n equations

$$y_i = a_0 + a_1 \cos 2\pi \frac{i}{n} + b_1 \sin 2\pi \frac{i}{n}; \; i = 0, 1, 2, \cdots, n-1, \qquad (4)$$

from which to determine a_0, a_1, b_1. If n exceeds 3, it is unlikely that all the equations (4) will be consistent; and the coefficients must then be found by least squares.

To this end we construct the normal equations by the usual method,[1] and have

$$a_0 = \frac{1}{n}\sum_{i=0}^{n-1} y_i, \; a_1 = \frac{2}{n}\sum_{i=0}^{n-1} y_i \cos 2\pi \frac{i}{n}, \; b_1 = \frac{2}{n}\sum_{i=0}^{n-1} y_i \sin 2\pi \frac{i}{n}. \qquad (5)$$

By the use of (5), one can calculate at once the coefficients of the function that fits the given series; and such a computation, for the data of Table I, is shown in Table II.

Simplified calculation. In practice the computation can be reduced considerably, because of certain symmetric properties of the sine and cosine. In general these schemes postpone the multiplication of the y_i by the various sine and cosine factors until certain necessary combinations have been made among the y_i themselves. The fundamental method, as developed by Runge,[2] applies if n is a multiple of 12; and a full description of this scheme is presented by Carse and Shearer.[3] An adaptation of the method to cases in which n is a multiple of 6 is given by Running,[4] and a scheme for n a multiple of 4 is to be found in Brunt's book.[5]

It should be remarked that, although these arithmetic schemes usually give equations for calculating the coefficients of all the harmonics as well as the terms of fundamental period, they are available for the limited case of three coefficients which we have in view. The point is, as we shall notice later, that each coefficient of a Fourier's series is independent of all the others: one can stop the calculation at any stage without needing to alter the coefficients already determined because of the omission of those not determined.

For other arithmetical methods of calculation and for certain graphical methods, reference is made to Carse and Shearer[6]; and, for the use of

[1] Brunt, *op. cit.*, pp. 77 *seq.* and 172 *seq.*

[2] *Zeitsch. für Math. und Phys.* (1903), p. 443; and (1905), p. 117.

[3] G. A. Carse and G. Shearer, *A Course in Fourier's Analysis and Periodogram Analysis* (London, Bell, 1915), chap. II.

[4] T. R. Running, *Empirical Formulas* (Wiley, 1917), chap. V.

[5] Brunt, *op. cit.*, pp. 181 *seq.*

[6] Carse and Shearer, *op. cit.*, chap. II.

harmonic analyzers, to the *Handbook*[1] of the Napier Tercentenary Celebration.

We show, in Table III, the abridged calculation by the method described by Running, for the series of Table I.

The analysis for multiple periods. Suppose next that y has more than one period, and that all the periods are sub-multiples of a single fundamental period. For the criticism and theoretical justification of the methods to be used, the reader is referred to a work on function theory[2] and to a recent report by Jackson[3]; but here we shall assume that y can be expressed by

$$\begin{aligned} y &= a_0 + a_1 \cos \frac{2\pi}{p} t + a_2 \cos 2 \frac{2\pi}{p} t + \cdots + a_r \cos r \frac{2\pi}{p} t \\ &\quad + b_1 \sin \frac{2\pi}{p} t + b_2 \sin 2 \frac{2\pi}{p} t + \cdots + b_r \sin r \frac{2\pi}{p} t \\ &= a_0 + \sum_{q=1}^{r} \left(a_q \cos q \frac{2\pi}{p} t + b_q \sin q \frac{2\pi}{p} t \right). \end{aligned} \tag{6}$$

If there are n observations and $n = 2r + 1$, the coefficients, a_0, a_q, b_q, can be determined exactly. If n exceeds $2r + 1$, we resort to the method of least squares and find for the normal equations

$$a_0 = \frac{1}{n} \sum_{i=0}^{n-1} y_i.$$

$$a_q = \frac{2}{n} \sum_{i=0}^{n-1} y_i \cos q\, 2\pi \frac{i}{n}, \quad b_q = \frac{2}{n} \sum_{i=0}^{n-1} y_i \sin q\, 2\pi \frac{i}{n}. \tag{7}$$

$$q = 1, 2, \cdots r.$$

These formulas lend themselves at once to calculation, and the various special computation schemes mentioned above are available. As already noted, the values of the various a_q and b_q are independent of each other; and the series may be broken off at any point. It can be shown,[4] however, that the larger we take r, up to the point where $n = 2r + 1$, the better will be the fit to the actual data.

The case for arbitrary functions. If the number of known values of y becomes infinite, which happens when we seek to express a known function by a Fourier's series, we may let r in (6) become infinite; and (6)

[1] *Modern Instruments and Methods of Calculation* (London, Bell, 1914).

[2] J. Pierpont, *Theory of Functions of Real Variables*, II (Ginn, 1912), chap. XIII.

[3] *Bull. Amer. Math. Soc.*, vol. 27 (1921), p. 415.

[4] Brunt, *op. cit.*, p. 177.

then takes the general form of Fourier's series. Introduce in (7) a new variable x

$$x = p\frac{i}{n}, \quad dx = \frac{p}{n}$$

and let n become infinite. The normal equations take the form

$$a_0 = \frac{1}{p}\int_0^p ydx, \quad a_q = \frac{2}{p}\int_0^p y \cos q\frac{2\pi}{p}xdx, \quad b_q = \frac{2}{p}\int_0^p y \sin q\frac{2\pi}{p}xdx \qquad (8)$$

We use (8) to calculate the coefficients of a Fourier's series to fit any arbitrary function. A graphical illustration of the approximations to the function $y = c$, where c is a ^onstant, is shown in Figure 16 for $r = 1, 2, 3, 4$.

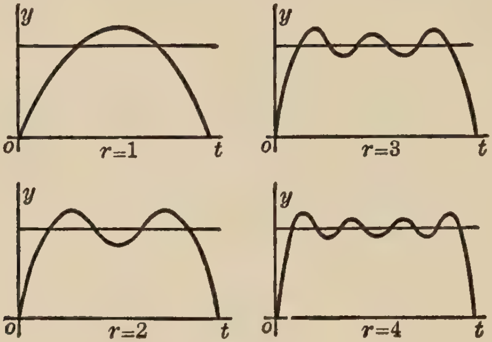


FIG. 16

THE PERIODOGRAM

Introductory discussion of the periodogram. If, instead of n observations at equal intervals p/n, there are kn observations extending over k periods, an obvious modification is made in the process. The values of y_i are arranged in rows, each containing n successive values of y_i, and each of the n columns is summed and the results are divided by k. These average values of y_i can then be treated by the method outlined above.

The method just described can be used if the statistical series is truly periodic and if the period is known. The fact is, however, that the "period" of most statistical time series is not exactly constant; and the direct use of the Fourier analysis may lead to faulty interpretations. Moreover, the length of the period may not be obvious from the series or its diagram, and it becomes necessary to devise a plan for finding periods. Finally, a Fourier's series, consisting of a term of fundamental period and its various harmonics, does not properly represent such series as exhibit the joint effect of several periodic terms for which periods are not commensurable. To attack these problems, we rely mainly upon the periodogram devised by Schuster.[1]

The validity of the periodogram analysis rests upon the fact that the summation method used tends to intensify the variation having a particular period, and to smooth out the other variations. To show the way this takes place we use an illustration adapted from Carse and Shearer.[2] Suppose we have, for successive instants of time, a series consisting of the sums of two periodic components:

$$0, 2, 4, 6, \ \ 7, \ \ 7, 6, 4, 2, 0, 2, 4, \cdots$$
$$0, 1, 2, 3, \ \ 4, \ \ 3, 2, 1, 0, 1, 2, 3, \ \ 4, \cdots$$

giving

$$0, 3, 6, 9, 11, 10, 8, 5, 2, 1, 4, 7, 10, 10, 9, 7, \cdots$$

If we arrange the first 63 terms of the resultant series in 7 rows of 9 terms each and add the columns, we get

$$15, 30, 47, 56, 62, 61, 55, 42, 29.$$

These values are given approximately by

$$7\left(\frac{0+1+2+3+4+3+2+1}{8}+i\right); \ i = 0, 2, 4, 6, 7, 7, 6, 4, 2, \quad (9)$$

and it appears therefore that the summation has served to average out the shorter period component and to multiply the effect of the longer period component by 7, the number of rows. This is due to the fact that the longer period component exactly completes a period in each row; whereas the shorter period component, which does not have an integral number of periods per row, suffers an advance in phase from row to row. It happens in this illustration that the two components are again in phase at the beginning of the ninth row; and, had we taken eight (or any multiple of eight) rows in computing the sums, we should have found that a rule such as (9) would fit the results exactly. For a number of rows not a multiple of eight the rule does not fit exactly. For instance, for eleven rows, the sums are

$$19, 44, 73, 92, 102, 99, 85, 62, 41$$

and the rule gives

$$22, 44, 66, 88, 99, 99, 88, 66, 44.$$

Had we taken nineteen rows, instead of eleven, the deviations from the rule would have been found numerically the same; but they would

[1] *Terrestrial Magnetism* (1898), pp. 13 *seq.*
Proc. Roy. Soc., A, vol. 77 (London, 1906), p. 136.

[2] Carse and Shearer, *op. cit.*, p. 30.

be less important relatively. The really significant fact is that long series of data must be used in periodogram work, if the single period under investigation is to be isolated.

If the rows had been taken with ten terms each, instead of nine, it is obvious that the summing process would have tended to smooth out the long-period component also. The essence of the periodogram method is to take that trial period, the length of a row, which will bring out most prominently the effect of the period being sought.

Changes effected by summing. Suppose the periodic component studied is

$$a \sin\left(\frac{2\pi}{p} t + \alpha\right)$$

and that we arrange our observations in n rows of m items each. The result of summing can be shown[1] to be

$$a \frac{\sin \dfrac{nm\pi}{p}}{\sin \dfrac{m\pi}{p}} \sin \left\{\frac{2\pi}{p} t + \alpha + (n-1)\frac{m\pi}{p}\right\}. \qquad (10)$$

We observe that the phase has been increased by $(n-1)\dfrac{m\pi}{p}$, and the amplitude has been multiplied by z, where

$$z = \frac{\sin n\pi x}{\sin \pi x}, \quad x = m/p$$

The maximum value of z is n, and occurs for $x = 1$: that is, for $m = p$. The value of z falls off rapidly on each side of $x = 1$, and becomes 0 for $x = 1 \pm 1/n$: that is, for $m = p\dfrac{n \pm 1}{n}$. For values of x beyond these points there are several maxima and minima of the z curve; but they are all relatively unimportant, as may be seen from Figure 17. It is then evident that, if we choose the right trial period, the summing process will yield a series of sums possessing a periodic variation of large amplitude; and, if our trial period deviates from the true period, the resulting amplitude will be much smaller.

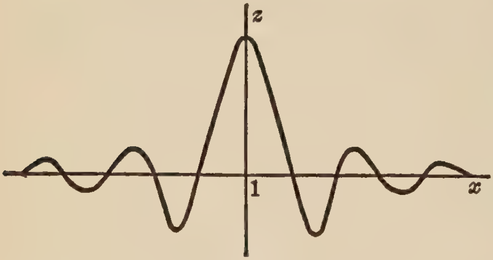


FIG. 17

[1] Carse and Shearer, *op. cit.*, p. 32.

Determination of a period. The process indicated by our discussion consists in selecting a trial period, m, arranging the data in rows of m items each, summing the columns, and computing the amplitude of the resultant series of sums by the Fourier method. The square of half this amplitude, multiplied by the total time interval covered by the statistics, is defined by Schuster [1] as the ordinate of the periodogram for the trial period m. Then another trial period, m', is taken, and an ordinate calculated for it. Schuster [2] has shown that two trial periods may be separated by an interval such that

$$m' - m = \frac{1}{4}\frac{\overline{m}}{N},$$

where $\overline{m}$ is approximately equal to m and m', and N is the total number of periods $\overline{m}$ in the whole time interval under investigation. It should be noted further, in order to agree with Schuster's analysis, that we assume here the time interval between two successive observations is unity.

When the ordinates have been calculated for a succession of trial periods m, we can plot them against the values of m as abscissas, and have then the periodogram. Values of m which yield large ordinates on this curve are in the vicinity of true periods, and they may be estimated from the curve itself. Owing, however, to the bluntness and lack of precision in the maxima of the periodogram, a careful determination of the period cannot be made from the curve. We use instead a scheme which is based upon the advance in phase suffered by a periodic term when the trial period differs from the true period, an advance noted in equation (10). If we have found a trial period m, which appears from the periodogram to be near a true period, we divide the n rows belonging to this trial period into groups of k successive rows each. It is then possible to calculate the average change in phase, β, from one group to the next, and the true period is given by [3]

$$p = \frac{m}{1 - \dfrac{\beta}{2\,\pi k}}.$$

The "oscillation" method. In order to save the labor of carrying through the Fourier analysis to get the amplitude for the sums for each trial period, an approximate value may be taken as one half the difference between the maximum and minimum of the averages obtained by dividing these sums by the number of rows. For a rough analysis,

[1] *Proc. Roy. Soc.*, A, vol. 77 (1906), p. 138.

[2] *Phil. Trans. Roy. Soc.*, A, vol. 206 (1906), p. 71.

[3] Brunt, *op. cit.*, p. 197.

these approximate values may be plotted at once as ordinates of the periodogram. Yule[1] has found that this method frequently leads to serious errors, and it must therefore be used with caution. The real difficulty seems to be due partly to the inherent dangers in using the method with statistical series which are too short or subject to frequent extreme deviations.

Other methods of finding the periodogram are discussed in the texts of Carse and Shearer, and of Brunt.

APPLICATIONS

Periodogram analysis of commercial paper rate. As an illustration of the application of the preliminary periodogram analysis, we shall examine the monthly record of the rate of interest on 60–90 day commercial paper in New York City from 1899 to 1914, as given in Table IV. We select this as an example because, although it is not very well fitted for periodogram analysis, the increasing use of the periodogram method in economic statistics renders an illustration from economic material particularly useful. A rough survey of the data indicates that, in addition to the period of twelve months corresponding to the marked seasonal variation, there is an approximate period of about 40 months. The portion of the periodogram in the vicinity of the 40-month trial period is shown in Figure 18, and the supporting data are given in Table VI. The detail calculations for the 43-month trial period are presented in Table V.

It is apparent that 43 months gives the greatest intensity and, on the basis of the preliminary study, that would be accepted as the period. It should be observed, however, that there are several ordinates in the vicinity of 43 which are nearly as large as that at 43. There are two chief reasons for this lack of precision in the periodogram. One is that the series is unduly short: only four trial periods as shown by Table V. It may be said at once that the use of the periodogram method on so short a series is quite unjustifiable except to give the approximate value of the period in order to assist in describing the fluctuations. The use of a period calculated from so short a series, to fit a periodic curve to the data, to substantiate a causal hypothesis, or to forecast the future, is extremely hazardous. Moreover, it is evident that the refinement of the result of the preliminary calculation, by the method outlined in Brunt (p. 197), is quite impossible in such an example. Likewise, it is not possible to break the series into halves and determine whether the same period

[1] *Jour. Roy. Stat. Soc.*, vol. 84 (1921), p. 525.

holds in both halves. Indeed, it cannot be too strongly emphasized that the full application of the periodogram method with its proper tests requires that the series under investigation be much longer than that of Table IV.

A second cause of the blurred maximum in the periodogram in the vicinity of 43 months is found in the nature of economic fluctuation. A historical economic series is characterized in general by frequent extreme and irregular deviations. The periodogram method does not effectively average out these extreme items; and, as they quite frequently occur in or near a particular phase of the cycle, they have a marked effect on the appearance of the periodogram. Perhaps the only sure way to discover the part played by these extreme items is to plot the entire record, using the times as abscissas and the data as ordinates. Moreover, the strong seasonal movement in the series under investigation probably contributes to the confusion of the periodogram. If the series were much longer, this influence would doubtless be less important; but even then it can be shown that trial periods which are commensurable with the seasonal period will yield results in which the influence of the seasonal movement enters in an unknown degree. The best that can be said is that the periodogram analysis of an economic record must be supplemented by a most searching criticism in the light of the actual data.

Critical remarks on the periodogram method. We have stated several times that it is important that the original data cover a long time interval. It cannot be insisted too strongly that this method of analysis is not valid for a short statistical series. On the other hand, in many phenomena the period undergoes a change in length with the lapse of time; and this will result in misleading inferences, if the periodogram is used blindly. An instance of the extreme care with which the method must be applied will be found in Schuster's Sunspot paper,[1] in which he finds a notable change in period. An investigation of overlapping periods has been made by Trachtenberg [2] in the study of statistics of epidemics.

The practice of breaking up the total time interval into segments, and analyzing each separately, should be followed where possible. For this purpose, each *part* of the series must be long. It is only in this way, however, that we can hope to reveal those changes in amplitude and phase, and even in period, which seem to be characteristic of many historical series.

[1] *Phil. Trans. Roy. Soc.*, A, vol. 206 (1906), p. 69.

[2] *Jour. Roy. Stat. Soc.*, vol. 84 (1921), p. 578.

In seeking the possibility of spurious results due merely to chance, Schuster[1] studied the probability of accidental appearance of a given amplitude. He has developed a formula which gives a minimum limit to the length of a series of observations if we are to rely upon the determination of amplitudes which are as small as a given fraction of the mean value. This study must serve as a warning against the careless use of the delicate instrument furnished by the periodogram.

Maxima and minima by inspection. In general, we may say that it is unwise to lose sight of the original statistics. If these be plotted, against the time as abscissas, it may often be possible to pick out by inspection the various maxima and minima. It is suggested[2] that the corresponding period may be computed by the use of

$$T = \frac{6}{n(n^2-1)}\{(n-1)(t_n-t_1)+(n-3)(t_{n-1}-t_2)+(n-5)(t_{n-2}-t_3)+\cdots\},$$

a graduation formula which will be found derived in Tuttle.[3] This method of finding T must be used with caution; for, if there are several component periodic terms in the series, it is quite likely that we shall be unable to pick out the maxima and minima by inspection of the graph.

It seems, therefore, that the periodogram offers the best present method for studying periodicities; but it must be applied only to series which are sufficiently long, and its use should be accompanied by a careful supplementary study of the segments of the series.

[1] *Terrestrial Magnetism* (1898), p. 18.

[2] By Professor Allyn A. Young.

[3] L. Tuttle, *The Theory of Measurements* (Phila., 1916), p. 245.

TABLES AND PERIODOGRAM CHART

Table I

t	0	$p/12$	$2\,p/12$	$3\,p/12$	$4\,p/12$	$5\,p/12$	$6\,p/12$	$7\,p/12$	$8\,p/12$	$9\,p/12$	$10\,p/12$	$11\,p/12$
y	35.9	38.6	36.6	29.9	20.7	13.1	7.2	3.3	6.3	12.1	21.7	30.5

(If p is not $2\,\pi$, we may make a change of variable so that subsequent calculations can be carried through on the assumption that p is $2\,\pi$.)

Table II

$n = 12$

i	$360\,i/n$	$\cos 360\,i/n$	$\sin 360\,i/n$	y_i	$y_i \cos 360\,i/n$	$y_i \sin 360\,i/n$
0	0	1.	0.0	35.9	35.9	0.
1	30	.87	.5	38.6	31.3	19.3
2	60	.5	.87	36.6	18.3	31.8
3	90	0.0	1.	29.9	0.	29.9
4	120	− .5	.87	20.7	− 10.4	18.0
5	150	− .87	.5	13.1	− 11.4	6.6
6	180	− 1.	0.0	7.2	− 7.2	0.
7	210	− .87	− .5	3.3	− 2.9	− 1.7
8	240	− .5	− .87	6.3	− 3.2	− 5.5
9	270	0.0	− 1.	12.1	0.	− 12.1
10	300	.5	− .87	21.7	10.9	− 18.9
11	330	.87	− .5	30.5	26.5	− 15.3
	Totals			255.9	87.8	52.1

From which: $a_0 = 21.325$, $a_1 = 14.63$, $b_1 = 8.68$, by use of (5).

Table III

	35.9	38.6	36.6	29.9	20.7	13.1	7.2	y_i, i from 0 to 6
		30.5	21.7	12.1	6.3	3.3		i from 11 to 7
v_i, 0 to 6	35.9	69.1	58.3	42.0	27.0	16.4	7.2	Sums (S)
w_i, 1 to 5		8.1	14.9	17.8	14.4	9.8		Differences (D)

	35.9	69.1	58.3	42.0	v_i, 0 to 3;	8.1	14.9	17.8	w_i, 1 to 3
	7.2	16.4	27.0		6 to 4	9.8	14.4		5 to 4
p_i, 0 to 3	43.1	85.5	85.3	42.0	r_i, 1 to 3	17.9	29.3	17.8	
q_i, 0 to 2	28.7	52.7	31.3		s_1 and s_2	− 1.7	.5		
	43.1	85.5	p_0 and p_1			17.9	28.7		r_1 and q_0
	85.3	42.0	p_2 and p_3			17.8	31.2		r_3 and q_2
L_0 and L_1	128.4	127.5			t_1 and t_2	.1	− 2.6		

Which give: $a_0 = 21.325$, $a_1 = 14.98$, $b_1 = 8.51$ and, by (3), amplitude is 17.3, phase angle is 60° 20′.

Table IV. Interest Rates on 60–90 Day Commercial Paper[1]

Deviations from 5 per cent; in units of $\frac{1}{10}$ per cent

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1899									−2	1	4	9
1900	−2	−6	−1	−7	−13	−13	−10	−8	−6	1	−6	−3
1901	−9	−13	−13	−10	−11	−11	−8	−5	−1	−4	−3	−1
1902	−4	−10	−6	−5	−5	−6	−4	−2	6	9	7	10
1903	2	−1	5	2	−3	2	4	9	10	8	10	8
1904	−1	−2	−3	−8	−11	−14	−15	−12	−7	−6	−9	−7
1905	−10	−12	−11	−10	−10	−13	−9	−8	−3	−1	5	8
1906	1	00	3	4	3	3	5	10	16	13	13	13
1907	12	9	12	9	4	5	8	13	18	21	24	30
1908	16	1	6	−6	−11	−13	−13	−14	−11	−9	−10	−12
1909	−13	−15	−15	−15	−16	−18	−16	−10	−8	00	1	1
1910	−3	−6	−5	−3	−3	−2	4	4	5	6	5	−3
1911	−10	−9	−11	−13	−14	−13	−12	−8	−5	−7	−11	−4
1912	−11	−13	−8	−9	−8	−10	−5	00	6	9	7	10
1913	−1	−1	8	5	4	9	11	10	8	7	6	7
1914	−5	−12	−11	−13								

[1] W. L. Crum, *Rev. Econ. Stat.*, January, 1923.

TABLE V. SUMMATION FOR TRIAL PERIOD OF 43 MONTHS, INTERVAL FROM NOVEMBER, 1899, TO FEBRUARY, 1914

	4	9	−2	−6	−1	−7	−13	−13	−10	−8	−6	1	−6	−3	−9	−13	−13	−10	−11	−11	−8
	2	4	9	10	8	10	8	−1	−2	−3	−8	−11	−14	−15	−12	−7	−6	−9	−7	−10	−12
	12	9	12	9	4	5	8	13	18	21	24	30	16	1	6	−6	−11	−13	−13	−14	−11
	4	5	6	5	−3	−10	−9	−11	−13	−14	−13	−12	−8	−5	−7	−11	−4	−11	−13	−8	−9
T_i	22	27	25	18	8	−2	−6	−12	−7	−4	−3	8	−12	−22	−22	−37	−34	−43	−44	−43	−40
−5	−1	−4	−3	−1	−4	−10	−6	−5	−5	−6	−4	−2	6	9	7	10	2	−1	5	2	−3
−11	−10	−10	−13	−9	−8	−3	−1	5	8	1	0	3	4	3	3	5	10	16	13	13	13
−9	−10	−12	−13	−15	−15	−15	−16	−18	−16	−10	−8	0	1	1	−3	−6	−5	−3	−3	−2	4
−8	−10	−5	0	6	9	7	10	−1	−1	8	5	4	9	11	10	8	7	6	7	−5	−12
−33	−31	−31	−29	−19	−18	−21	−13	−19	−14	−7	−7	5	20	24	17	17	14	18	22	8	2

$$\sum_{i=0}^{42} T_i \cos\frac{2\pi i}{43} = 675.$$

$$\sum_{i=0}^{42} T_i \sin\frac{2\pi i}{43} = -176.$$

$$A = \tfrac{2}{43}(675).$$

$$B = \tfrac{2}{43}(-176).$$

$$R = \sqrt{A^2 + B^2} = 32.5.$$

Table VI

Trial Period	Total Interval		R	Ordinate
	From	To		10,000
38	Sept. '00	Apr. '13	22.9	1.99
39	July '00	June '13	25.0	2.44
40	May '00	Aug. '13	26.5	2.81
41	Mar. '00	Oct. '13	27.7	3.15
42	Jan. '00	Dec. '13	28.3	3.36
43	Nov. '99	Feb. '14	32.5	4.54
44	Sept. '99	Apr. '14	25.9	2.95

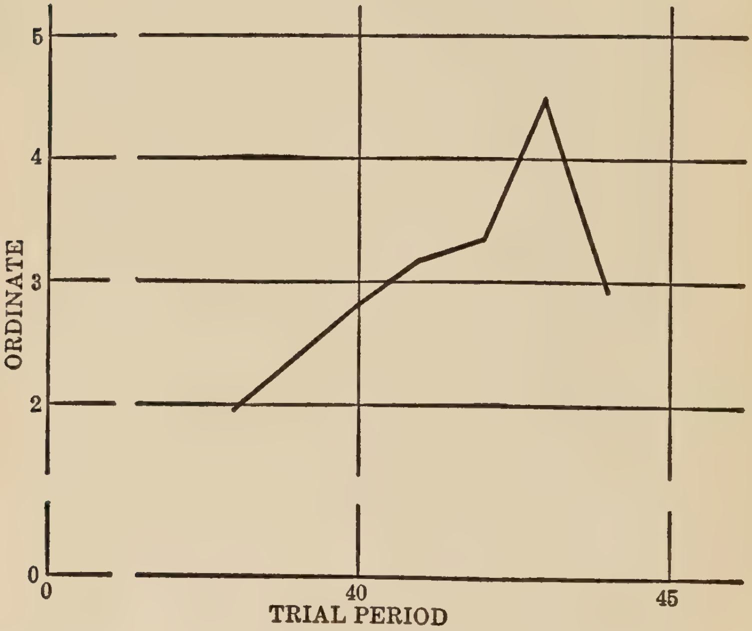


Fig. 18

CHAPTER XII

INDEX NUMBERS

By ALLYN A. YOUNG

Index numbers are series of numbers which measure and express the relative changes, as from time to time or from place to place, in the magnitude of statistical groups or aggregates of variables. Sometimes a series of numbers proportional to some other simple statistical series of numbers, not of groups, is called a series of index numbers. It is better, however, to refer to these simpler proportional series as series of **relative numbers** or merely as series of **relatives.**

In constructing a series of relatives some one term of the statistical series is selected as the **base**, its relative made 100 or unity, and the other relatives expressed as per cents or hundredths. Relative numbers have a large and growing use as an expository device, especially in economic statistics.

Except for the choice of the base, no special problems attach to the construction of series of relatives. Index numbers of group changes, however, present serious difficulties. Such index numbers are of three general types, corresponding to three different ways of expressing the magnitude of group change: (1) The changes undergone by each separate variable in the group may be expressed by a series of relatives, and the averages of such relatives taken to express the changes of the group. (2) The group may be represented by an appropriate average and a series of such averages expressed by relatives. (3) The individual variables in the group may be combined so as to form an aggregate, and a series of such aggregates expressed by relatives. Index numbers of the first type are **averages of relatives**; of the second type, **ratios of averages**; of the third type, **ratios of aggregates.**

I

AVERAGES OF RELATIVES

The arithmetic average. Index numbers have found their most important use in measuring price changes. The types most commonly used in the past for that purpose have been series of *arithmetic averages*, weighted or unweighted, of *relatives* which express the periodic changes of the different prices in a given group. It has been found, however, that unless they are weighted so as virtually to make them no longer averages of relatives, such index numbers are generally untrustworthy, and in particular that they are subject to biased errors.

If p_1, p'_1, p''_1, etc., are the prices of n different commodities in a year selected as a base, and p_2, p'_2, p''_2, etc., their prices in a second year, the unweighted arithmetic average of relatives for the second year, taking the base as unity, is

$$\frac{\frac{p_2}{p_1} + \frac{p'_2}{p'_1} + \frac{p''_2}{p''_1} + \cdots}{n}. \tag{1}$$

Let $p_2 = p_1 + \Delta p_1$, $p'_2 = p'_1 + \Delta p'_1$, etc., where Δp_1, $\Delta p'_1$, etc., may be positive or negative. Then the expression for the unweighted arithmetic average of relatives becomes

$$1 + \frac{\frac{1}{p_1}\Delta p_1 + \frac{1}{p'_1}\Delta p'_1 + \frac{1}{p''_1}\Delta p''_1 + \cdots}{n}.$$

The fundamental defect of the unweighted arithmetic average of relatives is that the weight given to the amount of change of the price of any commodity is inversely proportioned to the magnitude of the price of that commodity in the base year. As soon as prices change, dividing them by prices in the base year ceases to bring them to an exactly comparable basis — unless, indeed, all prices happen to change at the same rate. The greater the differences of the rates of change; that is, the greater their dispersion, the less accurately do the series of relatives represent them.

Biased error. In a series of increasing prices, p_1, if it be the first price in the series, will be *relatively* small. In a series of decreasing prices it will be relatively large. Adding the relatives computed on such a base, therefore, gives too much weight to increasing and too little to decreasing prices. Suppose, for example, that we are concerned with two commodities only. In the base year the price of one is \$1, of the other, \$2. In another year the first price becomes \$2, while the second falls to \$1. Giving equal weights to the two commodities, there is, by any reasonable test, no net movement of prices. Yet the price relatives become 200 and 50, and their arithmetic mean, 125, indicates an average increase in prices of 25 per cent. If we take the same figures and use the later year as a base, an average *fall* of prices from 125 to 100 — that is, of 20 per cent — is indicated. It should also be noted that the arithmetic average of relatives gives relatively more weight to prices that rise rapidly than to prices that rise slowly, and relatively more weight to prices that fall slowly than to prices that fall rapidly. This average, therefore, exaggerates a general rise and understates a general fall in prices. A purely random set of changes would be reported by the arithmetic average of relatives as an average increase.

Types of bases. In computing index numbers the base is sometimes *broadened* to include a number of years, or other periods. It may even be coextensive with the whole series of index numbers. Thus, for a series of relatives covering m years the base might be the arithmetic average, $\frac{p_1 + p_2 + \cdots + p_m}{m}$. The index number for any year k would then be

$$\frac{m}{n}\left(\frac{p_k}{\Sigma(p)} + \frac{p'_k}{\Sigma(p')} + \frac{p''_k}{\Sigma(p'')} + \cdots\right). \tag{2}$$

As compared with the single base taken at the beginning of the series a broadened base of the inclusive type, if the averages of both of the successive prices and of the relatives are arithmetic, will give a bias in the same direction if prices in general are decreasing, but will give an opposite bias if they are increasing. In either case it lessens somewhat the *amount* of biased error. Using as a base an average of prices in a relatively small group of successive years at the beginning of a relatively long series has a negligible effect in lessening bias. Its principal advantage is that it diminishes the chance that the base figures for one or more series of relatives will be distinctly abnormal or unrepresentative.

A *moving base* is sometimes utilized. In such case chain or **link relatives,** such as $\frac{p_2}{p_1}, \frac{p_3}{p_2}, \frac{p_4}{p_3}$, etc., take the place of fixed base relatives, such as $\frac{p_2}{p_1}, \frac{p_3}{p_1}, \frac{p_4}{p_1}$, etc. A series such as $\frac{\Sigma\left(\frac{p_2}{p_1}\right)}{n}, \frac{\Sigma\left(\frac{p_3}{p_2}\right)}{n}, \frac{\Sigma\left(\frac{p_4}{p_3}\right)}{n}$, etc., is called a series of **chain index numbers.**

The numbers in such a series may be referred to a fixed base by successive multiplications, so that the chain-derived fixed-base index number for the mth year from the base becomes:

$$\frac{\Sigma\left(\frac{p_2}{p_1}\right) \times \Sigma\left(\frac{p_3}{p_2}\right) \times \cdots \times \Sigma\left(\frac{p_m}{p_{m-1}}\right)}{n^{m-1}}. \tag{3}$$

No particular advantages have commonly been claimed for the chain-derived index number as a means of making comparisons with a fixed base. It is sometimes supposed, however, that chain index numbers derived from link relatives give more accurate direct *year-with-year* comparisons than do chain index numbers derived from arithmetic averages of fixed-base relatives by division, as, for example, for the mth year: $\Sigma\left(\frac{p_m}{p_1}\right) \div \Sigma\left(\frac{p_{m-1}}{p_1}\right)$. But the reverse is true.

In general year-with-year comparisons derived from arithmetic averages of fixed-base relatives are more trustworthy than the direct comparisons with the base which such averages give. The reasons are as follows: (1) Let M be the arithmetic average of such relatives. Dividing M_m by M_{m-1} eliminates all systematic or biased error, except for the excess of the error of one average over the error of the other. Even if the amount of systematic error increases proportionately or more than proportionately to the distance from the base, such excess or difference will in general be relatively small as compared with the error of either M_m or M_{m-1}. (2) In fact, however, the rate of increase of the systematic error of the average diminishes as its distance from the base increases. Centripetal forces in the price system tend to keep the trends of different prices in alignment. In any year prices that have already increased at more than the average rate are more likely to fall and less likely to rise than prices which have increased less than the average. Just so far as these tendencies operate, they lessen the disparity between the true rates at which prices change in the given year and the changes of the fixed-base relatives. Link relatives, however, start afresh each year, so that the disparities which create biased error are at their worst.[1]

The difference between the increase of prices between the year k and the year $k + 1$ as reported by the arithmetic average of link relatives and by the quotient of the arithmetic averages of fixed-base relatives for the years k and $k + 1$ is equal to $\frac{\Sigma(r_k d_k)}{n}$, when $(1 + r_k)$ is a link relative and d_k is the relative or percentage difference between a fixed-base relative, $\frac{p_k}{p_1}$, and the arithmetic average of such relatives in the year k.[2]

The only substantial advantage in the use of link relatives is that they permit of "splicing"; that is, of changing some of the series of

[1] F. R. Macauley, in *The American Economic Review*, vol. 6 (1916), p. 208, noted that "chain numbers draw away (upwards) from the fixed-base numbers," and rightly attributed it to "a greater tendency to rise and a less tendency to fall (in percentages) with the smaller relatives than with the larger relatives."

[2] This is, in slightly modified form, a result reached by W. F. Ogburn, in *Bulletin of the U.S. Bureau of Labor Statistics*, No. 284 (1921), p. 88.

price quotations used or even of increasing their number in accordance with changes in the available data. But certain other types of index numbers share this advantage, and some of them are preferable on other grounds to arithmetic averages of link relatives.

If the arithmetic average of relatives is to be employed for the direct comparison of prices in two different years or places, its bias may be eliminated by using the mean proportionals between prices at the two periods or places as bases. This gives the formula:

$$\frac{\Sigma\left(\frac{p_k}{\sqrt{p_1 p_k}}\right)}{n} \div \frac{\Sigma\left(\frac{p_1}{\sqrt{p_1 p_k}}\right)}{n}, \text{ or } \Sigma\sqrt{\frac{p_k}{p_1}} \div \Sigma\sqrt{\frac{p_1}{p_k}}. \qquad (4)$$

This is probably, in principle, the best form of the unweighted arithmetic average of relatives. It even has some points of superiority over the geometric average, with which in general it will agree very closely. If only two series of prices are used, it will be identical with the geometric mean.

The harmonic average. The unweighted harmonic average of relatives for the year *k* is:

$$\frac{n}{\frac{1}{\frac{p_k}{p_1}} + \frac{1}{\frac{p'_k}{p'_1}} + \frac{1}{\frac{p''_k}{p''_1}} + \cdots}, \text{ or } \frac{n}{\Sigma\left(\frac{p_1}{p_k}\right)}. \qquad (5)$$

It has sometimes been urged that the harmonic average is the proper measure of a change of the *purchasing power* of money, since purchasing power is the reciprocal of price, and a change in price is accompanied by an inverse change of purchasing power. The point is not sound. A change in price *is* an inverse change in purchasing power. In a properly constructed index number it should be a matter of indifference whether price changes or their reciprocals are the component units. But the harmonic average does not agree with the arithmetic average. It uniformly gives a smaller result. A glance at its formula will show that the harmonic average of relatives is the reciprocal of the arithmetic average with the base shifted to the other of the two years involved in the comparison. It has the same type of biased error as the arithmetic average, but its error is in the opposite direction. In general, therefore what has been said of the arithmetic average of relatives holds, *mutatis mutandis,* for the harmonic average. For example, just as the arithmetic average exaggerates a general rise of prices, so the harmonic average understates it.

Compromise index numbers. Because the biased errors of the arithmetic and harmonic averages of relatives are opposite in direction and similar in type, it is obvious that they may be eliminated, in large measure at least, by blending the two averages. In principle it is better to use their geometric rather than their arithmetic mean, on the ground that the relative rather than the absolute amounts of the biased errors are more nearly compensating. In practice, however, there is little difference between the formulas

$$\sqrt{\Sigma\left(\frac{p_k}{p_1}\right) \div \Sigma\left(\frac{p_1}{p_k}\right)},\ (6\,a)\ \text{and}\ \frac{1}{2}\left[\frac{\Sigma\left(\frac{p_k}{p_1}\right)}{n} + \frac{n}{\Sigma\left(\frac{p_1}{p_k}\right)}\right]. \qquad (6\,b)$$

This method of "rectification," or logical compromise, is applicable to a wide range of index numbers, weighted and unweighted, that exhibit systematic biased error.[1] Although unsatisfactory from the point of view of principle, and, in particular, complicated and difficult to compute, these compromise or "rectified" index numbers give, in general, satisfactory results. In practice it will be found that index numbers computed by formulas (4), (6 *a*), (6 *b*), and (7) will not differ notably.[2]

The geometric average. The unweighted geometric average has certain qualities that are highly desirable in index numbers. Some of these qualities are exhibited by the following elementary relations:

$$\sqrt[n]{\frac{p_k}{p_1} \times \frac{p'_k}{p'_1} \times \cdots (n \text{ terms})} = \frac{1}{\sqrt[n]{\frac{p_1}{p_k} \times \frac{p'_1}{p'_k} \times \cdots}} = \frac{\sqrt[n]{p_k \times p'_k \times \cdots}}{\sqrt[n]{p_1 \times p'_1 \times \cdots}}. \qquad (7)$$

The geometric average of relatives is the reciprocal of the geometric average of relatives with the base reversed. That is, it is *independent of the base.* Index numbers computed on one base can be shifted to another base by simple division. Splicing is therefore feasible. Chain and fixed-base methods give identical results. Moreover there is no difference between the unweighted geometric averages of relatives and the ratios between geometric averages of actual prices. In brief, geometric averages give index numbers that are *self-consistent.* Furthermore, the geometric average is, on logical grounds, an appropriate average

[1] The possibilities of this method have been systematically explored by Irving Fisher, in *The Making of Index Numbers* (1922). The method is not applicable when the two series are both either arithmetic or harmonic averages.

[2] The relation between the geometric mean of the arithmetic and harmonic averages and the geometric average (formulas 6 *a* and 7) has been investigated by C. M. Walsh, in *The Measurement of General Exchange-Value* (1901), pp. 516–19.

of relatives. It has no systematic bias. It is a true average (although not the only possible type of true average) of *rates of change.*

The median. Simplicity and definiteness of meaning are the principal advantages of the median. It has no biased error. But it is unreliable and erratic unless the number of prices (or other items) in the group is fairly large. It is insensitive to forces that show themselves in price changes only on one side or the other of the median. Nevertheless medians of price relatives generally give much more reliable unweighted index numbers than do arithmetic averages, other than the special type (4).

RATIOS OF AVERAGES AND OF AGGREGATES

Ratios of averages. Arithmetic averages of actual unit prices[1] do not have the biased errors that inhere in arithmetic averages of relatives. But they suffer from the fact that unit prices of different goods differ greatly in magnitude, varying not only with the value of the commodity, but also with the physical unit (for example, ton, ounce, yard) the price of which is quoted. The " unweighted " average is in reality arbitrarily weighted.[2] The harmonic average, utilizing the reciprocals of unit prices, reverses the weighting. In the one case single units of goods and their money prices are the components of the average; in the other case units of money (dollars) and " dollar's worths " of goods. The difficulty may be removed, however, and the arithmetic and harmonic averages brought into agreement by weighting the arithmetic average proportionately to physical quantities of goods (q) and the harmonic by quantities of money, or values (qp), for

$$\frac{\Sigma(qp)}{\Sigma(q)} = \frac{\Sigma(qp)}{\Sigma\left(qp\frac{1}{p}\right)}. \tag{8}$$

A *series* of index numbers such as $\frac{\Sigma(q_1p_1)}{\Sigma(q_1)}$, $\frac{\Sigma(q_kp_k)}{\Sigma(q_k)}$, etc., however, has the defect that if the weights of commodities whose unit prices happen to be large increase faster than the other weights, the index number will be too large, and *vice versa.* This source of error may be removed by converting the weights into " dollar's worths." When the conversion

[1] For a more detailed analysis see A. A. Young, "The Measurement of Changes of the General Price Level," *Quarterly Journal of Economics*, vol. 35 (1921), pp. 557–73.

[2] Reducing physical units to pounds (as in Bradstreet's index numbers) or to some other common measure merely introduces another sort of arbitrary weighting.

is made on the basis of prices in the first year, the index number for the other year becomes the ratio of aggregates,

$$\frac{\Sigma(q_k p_k)}{\Sigma(q_k p_1)}. \tag{9}$$

Converting on the basis of prices in the second year gives for that year

$$\frac{\Sigma(q_1 p_k)}{\Sigma(q_1 p_1)}. \tag{10}$$

Ratios of aggregates. Ratios of aggregates have long been recognized as appropriate expressions of the changes of " composite prices," such as the cost of living. There has been growing recognition of their merits as measures of general price movements. Analytically, as we have seen, they are akin to ratios of averages, and for that reason the two types are here considered together.

The two fundamental forms are (9) and (10) above. There is no general ground for preferring one to the other. A compromise is therefore logically indicated, such as

$$\sqrt{\frac{\Sigma(q_k p_k)}{\Sigma(q_k p_1)} \times \frac{\Sigma(q_1 p_k)}{\Sigma(q_1 p_1)}}. \tag{11}$$

Professor Irving Fisher holds, on weighty grounds, that formula (11) is the " ideal " index number. He has shown that it gives results intermediate between those given by forms which, compared with each other, have opposed types of biased error, and that, in general, the more trustworthy an index-number formula is, the more closely its results approximate those given by formula (11).[1]

A compromise between formulas (9) and (10) may also be effected by using geometric means of the weights rather than of the formulas themselves. This procedure gives:

$$\frac{\Sigma\sqrt{q_1 q_k}\, p_k}{\Sigma\sqrt{q_1 q_k}\, p_1}. \qquad (12)^2$$

As (9) and (10) are usually not far apart, substituting arithmetic for geometric means in (11) and (12) has a negligible effect upon their accuracy and makes them easier to compute. Formula (12) then becomes

$$\frac{\Sigma[\frac{1}{2}(q_1 + q_k)p_k]}{\Sigma[\frac{1}{2}(q_1 + q_k)p_1]}, \text{ or } \frac{\Sigma[(q_1 + q_k)p_k]}{\Sigma[(q_1 + q_k)p_1]}. \tag{13}$$

[1] *The Making of Index Numbers, passim.*

[2] This is identical with (4) — the best of the arithmetic averages of relatives — weighted by the geometric means of the money values of quantities of goods at the two periods; that is, by $\sqrt{p_1 q_1 \times p_k q_k}$.

When accuracy, simplicity, and ease of computation are all taken into account, this appears to afford as good year-with-year comparisons as any other single index number of prices.[1] Index numbers like (11) and (13) encounter the practical difficulty, however, that (except in the study of special fields, such as the prices of agricultural products) reliable figures for even approximately complete annual, to say nothing of monthly, weights are lacking.

Another practical difficulty with compromise forms like (11) and (13) is that although they afford probably the most accurate year-with-year comparisons, they are not so well adapted to the constructions of series of successive index numbers.[2] Chain index numbers, in which the weights and prices used are those of contiguous years, and fixed-base index numbers, computed by these formulas, will not agree. The fixed-base method is to be preferred by reason of the way in which the chain method accumulates error. But under most conditions the simple aggregative with fixed weights (10) is to be preferred when a self-consistent series, rather than year-by-year comparisons, or comparisons of successive years with the basing year, is the desideratum. Such numbers, however, should be checked from time to time by (9), (11), or (13) and, if necessary, their weights revised. Moreover, there is little difference between the accuracy of aggregative and geometric types of index numbers. Peculiarities of the available data may indicate that, in some particular use, a weighted geometric average should be preferred.

When weights are not available ratios of aggregates are not trustworthy, and recourse must be had to other methods. Although there is no large difference in the results given by (4), (6 *b*), and (7), its other advantages, and especially the fact that it is self-consistent (that is, independent of the base), probably entitle the geometric average (7) to preference, even though (4) may be, in general, slightly more trustworthy for year-with-year comparisons.

[1] Professor Fisher ranks (13) practically as high as (11). It has also the weighty approval of Mr. C. M. Walsh and Professors Alfred Marshall and F. Y. Edgeworth. In practice it seems to agree more closely with (11) than does (12) and it is probably somewhat more accurate than (12). Cf. Fisher, *op. cit.*, pp. 401–07.

[2] Cf. the findings of Professor W. M. Persons, *Review of Economic Statistics*, Prel. vol. 2, pp. 112, 113 (May, 1921). The difficulty is not with the particular formulas, which are probably the best of their kind. No aggregative index number with changing weights can meet the so-called "circular test"; that is, the test of self-consistency.

METHODS OF WEIGHTING

The effect of weighting. It has sometimes been held that the weighting of averages of *relative* prices may be dispensed with, provided that the number of series of price quotations used is relatively large. This contention would be well founded if there were no correlation between the importance of a commodity and its tendency to rise or fall in price, and further, if price variations always fell within a fairly narrow range in respect to magnitude. It is the failure of the second rather than the first of these conditions that makes weighting desirable. The unweighted index number is too sensitive to abnormal variations of relatively unimportant prices and is too little influenced by large variations of important prices.

Professor Mitchell infers, from a study of standard index numbers, that, except in abnormal years, weighting seldom makes a difference of 10 per cent.[1] But this, as he suggests, is a much larger margin of error than is allowable in a good index number. Furthermore, the problem has significance only for averages of relatives. Ratios of aggregates are of necessity weighted.

But weighting need not be precise. Round sums or even rough estimates will often serve the purpose about as well as precise figures. Professor Fisher has suggested that weighting to the nearest power of 10, effected merely by moving the decimal point, is sufficiently accurate for most purposes. Nor is it always necessary to weight all of the series. Weighting should give their due importance, and no more, to *large variations*, whether these be abnormal variations of the prices of unimportant commodities, or smaller variations of the prices of important commodities.

What weights are best? Weights must be selected with reference, first, to the type of formula used and, second, to the purpose of the particular index number. Thus in aggregative index numbers of prices the weights are necessarily physical units. Aggregative index numbers of the cost of living should likewise be weighted by physical quantities. In practice the component units of such index numbers are often *ab initio* "sums expended"; that is, prices weighted by quantities.

The proper weighting of the geometric average is a more difficult

[1] *Bulletin of the U.S. Bureau of Labor Statistics*, No. 284, p. 60. His comparisons are not wholly satisfactory, however, for the system of weighting employed is such as to change the type of the index numbers he studies. But other evidence, including his own comparisons of Dun's, Bradstreet's, and the Bureau of Labor's index numbers, supports his conclusion.

problem. In practice, unless there is a marked positive or negative correlation of the p's and q's, it is generally best to take values (such as pq) as weights, so that the weighted geometric, with constant weights, becomes

$$\sqrt[\Sigma pq]{\left(\frac{p_k}{p_1}\right)^{pq} \times \left(\frac{p'_k}{p'_1}\right)^{p'q'} \times \cdots (n \text{ terms})}. \qquad (14)$$

The same general principles of weighting hold in dealing with special types of prices, such as wages and interest rates. In index numbers of wages, however, constant weights give untrustworthy results, for increasing wages are likely to be correlated with increasing numbers of wage-earners.[1]

In constructing index numbers of the physical volume of trade, or of production or consumption, where the data run in terms of heterogeneous physical units, the choice of weights must be determined on different principles. When, as is often the case, there is no correlation between the importance of the different variables and their rates of change, and when the number of series is fairly large — say, twenty or more — an unweighted average of relatives, such as the median or the geometric (7), will generally give fairly reliable results. Otherwise different sorts of goods must be reduced to the only practicable common measure, money value — a procedure consistent with some uses of such index numbers, but not with all. Weighted index numbers may then be formed by any of the reliable aggregative or geometric formulas, such as (9), (10), (11), (13), (14), with the important difference that the p's and q's are interchanged.[2]

Just what should weights represent? Quantities (or values) consumed? or produced? or exchanged? For index numbers of the cost of living and of retail prices in general weighting by amounts consumed is indicated. For certain types of studies of the influence of changes of the quantity of money upon the price level, weighting of wholesale (and of retail) prices by amounts exchanged is desirable. But index numbers of wholesale prices probably serve the broadest range of interests, including the interests of economists, of business men, and, in general, of citizens in their dual rôles of producers and consumers, when they are weighted according to amounts produced.

Whatever the basis of weighting it is often desirable that commod-

[1] For other special problems of index numbers of wages see A. L. Bowley, *Elements of Statistics* (fourth ed., 1920), chap. IX, and references there given.

[2] In using the weighted geometric (14), "values" (such as pq) are probably the best practicable weights when the variables averaged are physical quantities. Cf. E. E. Day, "An Index of the Physical Volume of Production," *Review of Economic Statistics*, Prel. vol. 2 (Sept., 1920), p. 255.

ities should be classified in sub-groups, selected according to industries represented, or according as they are raw materials or finished products, or imported or exportable goods or goods sold mainly in the domestic market, or by other significant criteria. It is often quite as important to compare the fluctuations and trends of different groups of prices as to be informed of the movement of the general price level. The separate weighted indexes (if of the aggregative type) may generally be combined into weighted general index numbers without introducing error.[1]

Biased weighting. When sums of money or values (such as pq) are used as weights, care must be taken to avoid bias. Prices are factors in the weights, so that values for the later of any two periods compared are commonly directly correlated with price changes. The correlation between price changes and values of the earlier of any two periods compared is inverse. In the special case of the arithmetic average of relatives, however, weighting by basing-year values largely avoids bias, for

$$\frac{\Sigma\left(q_1 p_1 \frac{p_2}{p_1}\right)}{\Sigma(q_1 p_1)} = \frac{\Sigma(q_1 p_2)}{\Sigma(q_1 p_1)},$$

which is the aggregative formula (10).[2] Similarly, weighting the harmonic average of relatives by stated-year values (such as $q_2 p_2$) gives the other fundamental aggregative formula (9). But weighting the arithmetic average by stated-year values or weighting the harmonic average by basing-year values gives wholly unreliable results, for in each case the bias of the weighting reinforces the bias inherent in the unweighted form.

The geometric average, because its weights are exponents, is particularly sensitive to biased weighting. Using either basing-year or stated-year values as constant weights will introduce a considerable element of error, especially when there has been a marked general upward or downward movement of prices. The error may be lessened by using means (preferably geometric) of basing-year and stated-year values, or, for *series* of index numbers, averages (preferably geometric) of values for the years covered. Similar precautions are desirable in constructing weighted index numbers of physical production or of the physical volume of trade, by reason of the presence of q as a double factor.

In general, there is very little weight bias in ratios of aggregates, unless there is a marked degree of correlation between the p's and q's.

[1] For certain difficulties that may be encountered in practice, however, see Mitchell, *op. cit.*, p. 67.

[2] In this manner, it will be noted, ratios of aggregates are related to averages of relatives, as they are to ratios of averages.

In constructing index numbers of the cost of living, weighted in proportion to the relative importance of different types of consumable goods in family budgets, the use of a " crossed " formula,[1] such as (11) or a " crossed-weighted " formula, such as (13), is desirable, because of the inverse correlation commonly found between prices and quantities of goods consumed. For wholesale prices constant weights are less objectionable, because the correlation, positive or negative, of wholesale prices and quantities of goods exchanged is generally small. With a reasonably large number of series of price quotations formula (10), the simple aggregative with constant weights, gives trustworthy results over a fairly long period of years.[2]

THE ACCURACY OF INDEX NUMBERS

Substantial errors have often been put into index numbers by improper methods of construction. The errors inherent in the best formulas, however, are exceedingly small, as is shown by the close agreement of the results they give.

Like other averages, index numbers gain in accuracy when the number of constituent items is increased. The probable errors of sampling of unweighted averages of relatives may be computed by the ordinary rules. Professor Truman L. Kelley[3] suggests as a measure of the probable error of an index number of the aggregative type $.6745\,\sigma\sqrt{\frac{1-r}{2}}$ (r being the coefficient of correlation between the series of index numbers for two random halves into which the series of quotations is divided, and σ being the mean of the standard deviations of the two sub-series).

[1] Professor Irving Fisher suggests that weight bias may be eliminated by using the mean of a given formula and of a formula with an equal but opposed weight bias. This opposed formula, or "factor antithesis," is found by interchanging the p's and q's in the given formula, and dividing the result into the ratio,

$$\frac{\Sigma(q_k p_k)}{\Sigma(q_1 p_1)}.$$

Except in the case of ratios of aggregates, this method generally leads to cumbersome formulas. Results practically as good can be obtained by "crossing" (that is, by taking the mean of) weights rather than formulas.

[2] In 1922 the U.S. Bureau of Labor Statistics substituted weights based on data for 1919 for weights based on 1909 data in its index number of wholesale prices (327 commodities, formula 10). With the 1909 weighting the increase of prices from 1909 to 1919 was reported as 219 per cent. With the weights of 1919 the increase reported was larger, but by less then 3 per cent. Using formula (11) or (13) would have made a difference of less than 1.5 per cent. Cf. Fisher, *op. cit.*, p. 369.

[3] *Statistical Method* (1923), p. 338.

Professor Irving Fisher[1] has constructed index numbers of the prices of from 3 to 200 commodities equably apportioned, so far as possible, among the various distinctive classes. His results (percentage standard deviations) are as follows: 100 commodities, deviation 1.78; 50, 2.05; 25, 1.61; 12, 2.64; 6, 4.31; 3, 3.65.[2] He infers that "to reduce the error by half we must multiply the number of commodities not by four [as by the ordinary square-root rule] but by thirty-five." These results are largely accounted for, however, he suggests, by the fact that the least important commodities were discarded first. By extrapolating his results graphically he estimated the probable error of the index number of the complete sample of 200 commodities to be about 1.5 per cent. Professor Kelley's method gave 1.3 per cent.

"Seldom, however," Professor Fisher concludes, "are index numbers of much value unless they consist of more than 20 commodities; and 50 is a much better number. After 50, the improvement obtained from increasing the number of commodities is gradual and it is doubtful if the gain from increasing the number beyond 200 is ordinarily worth the extra trouble and expense."[3]

It is exceedingly important, however, that the quotations or other statistics used be a representative sample. More depends upon *what* commodities are included than upon their number. In constructing and using index numbers it is well to observe the effects of including or excluding some of the more heavily weighted or more variable items.

[1] *The Making of Index Numbers*, p. 337.

[2] Professor Mitchell's somewhat similar experiments with large and small index numbers (*Bulletin of the U.S. Bureau of Labor Statistics*, No. 284, pp. 34–41) lead him to put somewhat more emphasis on the difficulty of securing a representative sample in a small index number. The smaller the index number, it is clear, the greater is the care with which the commodities that enter into it must be selected.

[3] Professor W. M. Persons has found that much of the cyclical fluctuation, as distinguished from the trend, of index numbers of wholesale prices can be accounted for by the movements of a small number of prices. For use in forecasting business conditions, he has constructed a serviceable index number of the prices of only ten commodities. Cf. *Review of Economic Statistics*, November, 1921.

BIBLIOGRAPHY

BY W. L. CRUM

IN compiling a bibliography to accompany the *Handbook* no attempt has been made to present an exhaustive list of the writings, or even the more important writings, on the subject. The purpose has been rather to present, for the guidance of users of the *Handbook*, references to those treatises and memoirs which bear directly upon and amplify the methods developed in the text.

It is undoubtedly desirable that every title in the bibliography be followed by a synopsis and critical remarks. The task of providing such notes in any scientific subject is a burdensome one, and it is peculiarly difficult in the science of statistics because of the recent exceedingly rapid growth and consequent inadequate critical study of methodology. For the purposes of the *Handbook*, we have confined our summaries and critical remarks to certain treatises.

A large number of current scientific journals contain some papers on statistical methods but the following journals may be especially mentioned as devoted to the advancement of such methods:

Biometrika; Cambridge, England, The University Press.

Journal of the American Statistical Association; Concord, N. H., The Rumford Press. (Prior to June, 1922, the *Journal* was entitled the *Quarterly Publications of the American Statistical Association.*)

Journal of the Royal Statistical Society, London, published by the Society.

Metron; Padua, Tipografia Industrie Grafiche Italiane Padova.

Review of Economic Statistics; Cambridge, Harvard University Press.

In the bibliography the names of many journals are abbreviated, but it is believed that the abbreviations require no explanation except possibly as follows:

Biom. = *Biometrika.*

J. A. S. A. = *Journal of the American Statistical Association.*

Q. P. A. S. A. = *Quarterly Publications of the American Statistical Association.*

J. R. S. S. = *Journal of the Royal Statistical Society.*

Adams, T. S. Index Numbers and the Standard of Value. *Jour. Pol. Econ.*, 1901 and 1902.

Anderson, O. von. Nochmals über "The Elimination of Spurious Correlation due to Position in Time and Space." *Biom.*, vol. 10 (1914).

Bailey, W. B. *Modern Social Conditions.* N.Y., Century, 1906.

Bailey, W. B., and Cummings, J. *Statistics.* Boston, McClurg, 1917.

Barnett, G. E. Index Numbers of the Total Cost of Living. *Quar. Jour. Econ.*, vol. 35 (1921).

Benini, R. *Principii di statisticametodologica.* Torino, Unione tipografico — Editrice, 1906.

Bertillon, J. *Cours élémentaire de statistique.* Paris, Soc. d'éd. Sci., 1895.

Beveridge, Sir W. H. Wheat Prices and Rainfall in Western Europe. *J. R. S. S.*, vol. 85 (1922).

Blakeman, J. On Tests for Linearity of Regression in Frequency Distributions. *Biom.*, vol. 4 (1905).

Blakeman, J., and Pearson, K. On the Probable Error of the Coefficient of Mean Square Contingency. *Biom.*, vol. 5 (1906–07).

Blaschke, Ernest. *Vorlesungen über mathematische Statistik* (*die Lehre von den statistischen Masszahlen*). Leipzig, Teubner, 1906.

A treatise devoted largely to the theory of the flow of populations.

Block, Maurice. *Traité théorique et pratique de statistique.* Paris, Guillaumin, 1886.

Boole, G. *A Treatise on the Calculus of Finite Differences.* 3d ed., London, Macmillan, 1872.

Bortkiewicz, L. von. Anwendung der Wahrscheinlichkeitsrechnung auf Statistik. *Encyk. der math. Wiss.*, Bd. 1, H. 6; and *Encyc. der sci. math.*, tome 1, vol. 4, Fascicule 3.

—— *Das Gesetz der kleinen Zahlen.* Leipzig, Teubner, 1898.

—— Kritische Betrachtungen zur theoretischen Statistik. *Jahrb. für Nat. Ök. u. Stat.* (3), vol. 8 (1894).

Bowley, A. L. *An Elementary Manual of Statistics.* London, Macdonald Evans, 1910.

—— *Elements of Statistics.* London, P. S. King, 4th ed., 1920.

The work is presented in two parts, of which the first is in the main non-mathematical, and the second chiefly theoretical. The illustrations are drawn chiefly from economic subjects.

—— The Measurement of Changes in the Cost of Living. *J. R. S. S.*, vol. 82 (1919).

—— *The Measurement of Groups and Series.* London, C. & E. Layton, 1903.

Brinton, W. C. *Graphic Methods for Presenting Facts.* N.Y., Eng. Mag. Co., 1914.

Brown, W., and Thomson, G. H. *The Essentials of Mental Measurement.* Cambridge University Press, 1921.

Bruns, H. *Wahrscheinlichkeitsrechnung und Kollektivmasslehre.* Leipzig, Teubner, 1905.

A treatise devoted mainly to the representation of an arbitrary frequency function.

Brunt, David. *The Combination of Observations.* Cambridge University Press, 1917.

A valuable introductory treatment of the method of least squares. The first and chief part of the book is devoted to the theory and method of adjustment of observations, and is written in such a way as to render it useful both for the study of the foundations of the method and for the criticism of actual practice. It is not designed, however, as a laboratory manual of least squares. The later chapters give excellent summaries of the methods of curve-fitting, correlation, harmonic analysis, and periodogram analysis.

Buchanan, James. Osculatory Interpolation by Central Differences, with an Application to Life-Table Construction. *Jour. of the Inst. of Actuaries*, vol. 42 (1908).

Byerly, W. E. *Fourier's Series.* Boston, Ginn, 1893.

Carse, G. A., and Shearer, G. *A Course in Fourier's Analysis and Periodogram Analysis.* London, Bell, 1915.

Carver, H. C. The Mathematical Representation of Frequency Distributions. *Q. P. A. S. A.*, vol. 17 (1921).

Cave, B. M. *Cf.* Soper, H. E.

Cave, B. M., and Pearson, K. Numerical Illustrations of the Variate Difference Correlation Method. *Biom.*, vol. 10 (1915).

Charlier, C. V. L. Contributions to the Mathematical Theory of Statistics. *Arkiv. för matematik, astronomi och fysik,* vols. 7, 8, 9.

—— Researches into the Theory of Probability. Lund, *Contributions* from the Astronomical Observatory, 1906.

—— *Vorlesungen über die Grundzüge der mathematischen Statistik.* Lund, 1920.

Chauvenet, Wm. *A Manual of Spherical and Practical Astronomy.* Phila., Lippincott, 1863.

Crathorne, A. R. Calculation of the Correlation Ratio. *J. A. S. A.*, vol. 18 (1922).

—— Correlation Method Applied to Grades. *Report* of the National Committee on Mathematical Requirements, 1923.

Crofton, M. W. On the Proof of the Law of Errors of Observations. *Phil. Trans.*, 159 (1869), and "Probability," *Encycl. Brit.*, 19 (1885).

Crum, W. L. Cycles of Interest Rates on Commercial Paper. *Rev. Econ. Stat.*, vol. 5 (1923).

—— A Measure of Dispersion for Ordered Series. *Q. P. A. S. A.*, vol. 17 (1921).

—— A Special Application of Partial Correlation, *Q. P. A. S. A.*, vol. 17 (1921).

—— The Determination of Secular Trend. *J. A. S. A.*, vol. 18 (1922).

—— The Use of the Median in Determining Seasonal Variation. *J. A. S. A.*, vol. 18 (1923).

—— The Resemblance between the Ordinate of a Periodogram and the Correlation Coefficient. *J. A. S. A.*, vol. 18 (1923).

Cummings, J. *Cf.* Bailey, W. B.

Czuber, E. *Die statistichen Forschungs-methoden.* Wien, L. W. Seidel & Sohn, 1921.

This book resembles, in its methods, Yule's *Introduction to the Theory of Statistics.*

—— *Wahrscheinlichkeitsrechnung.* Leipzig, I, 3d ed., 1914; II, 3d ed., 1921.

Volume I, *Wahrscheinlichkeitstheorie, Fehlerausgleichung, Kollektivmasslehre.*

Volume II, *Mathematische Statistik, mathematische Grundlagen der Lebensversicherung.*

One of the leading treatises on statistical methods and error theory.

Darbishire, A. D. Some Tables for Illustrating Statistical Correlation. *Mem. & Proc. Manchester Lit. & Phil. Soc.*, vol. 51 (1907).

Davenport, C. B. *Statistical Methods, with Special Reference to Biological Variations.* 2d ed., N.Y., Wiley, 1904.

Day, E. E. The Physical Volume of Production. *Rev. Econ. Stat.*, vols. 2 and 3 (1920, 1921).

Dormoy, E. *Théorie mathématique des assurances sur la vie.* Paris, Gauthier-Villars, 1878.

Edgeworth, F. Y. Index Numbers. Palgrave's *Dictionary of Pol. Econ.*

—— "Law of Error" in 10th ed., and "Probability" in 11th ed., *Encyc. Brit.*

—— Miscellaneous Applications of the Calculus of Probabilities. *J. R. S. S.*, vols. 60, 61 (1897–98).

—— On Correlated Averages. *Phil. Mag.*, Series 5, vol. 34 (1892).

—— On Methods of Statistics. *J. R. S. S.*, Jubilee Volume (1885).

—— On the Method of Ascertaining a Change in the Value of Gold. *J. R. S. S.*, vol. 46 (1883).

—— On the Mathematical Representation of Statistical Data. *J. R. S. S.*, vols. 79, 80 (1916, 1917).

—— On the Method of Least Squares. *Phil. Mag.*, series 5, vol. 16 (1883).

—— On the Representation of Statistical Frequencies by a Series. *J. R. S. S.*, vol. 70 (1907).

—— On the Use of Analytic Geometry to Represent Certain Kinds of Statistics. *J. R. S. S.*, vol. 77 (1914).

—— Reports of the Committee (of the British Association for the Advancement of Science) appointed for the purpose of investigating the best methods of ascertaining and measuring variations in the value of the monetary standard. In *Reports of the Association*, 1888, 1889, 1890.

—— The Asymmetrical Probability Curve. *Phil. Mag.*, vol. 41 (1896).

—— The Generalized Law of Error or Law of Great Numbers. *J. R. S. S.*, vol. 69 (1906).

—— The Law of Error. *Camb. Phil. Trans.*, vol. 20 (1905).

Elderton, E. M. *Cf.* Elderton, W. P., and Pearson, K.

Elderton, W. P. *Frequency Curves and Correlation.* London, C. & E. Layton, 1906.

An exposition of the method of moments and the Pearson system of frequency curve-fitting. The final chapters treat of various methods of measuring correlation. Although the book is written with actuarial problems in view, it is readily understandable by the general student.

—— Graduation and Analysis of a Sickness Table. *Biom.*, vol. 2 (1902–03).

—— Interpolation by Finite Differences. *Biom.*, vol. 2 (1902–03).

—— Notes on Statistical Processes. *Biom.*, vol. 4 (1905–06).

—— Tables of Powers of Natural Numbers and of Sums of Powers of Natural Numbers from 1 to 100. *Biom.*, vol. 2 (1902–03).

Elderton, W. P., and E. M. *Primer of Statistics.* London, A. & C. Black, 1910.

Everett, J. D. On a New Interpolation Formula. *Jour. Inst. Actuaries*, vol. 35, 1901.

Fechner, G. T. *Kollektivmasslehre, herausgegeben von G. F. Lipps.* Leipzig, Engelman, 1897.

This book presents the work of Fechner on the problem of generalized frequency theory.

FECHNER, G. T. Über den Ausgangswerth der kleinsten Abweichungssumme, u. s. w., Leipzig, *Abh. der Kg. sächs. Gesell. der Wiss.*, vol. 18 (1878).

FILON, L. N. G. *Cf.* Pearson, K.

FISHER, ARNE, *The Mathematical Theory of Probabilities*, N.Y., Macmillan, 1922.

A careful development of the classical theory of probability, in the light of the work of recent scholars and in particular of the Scandinavian school led by Charlier. In Part II the author treats of frequency curves and heterograde statistics, and adopts here also the point of view of Charlier. The book is one of the most important recent works on the subject, and merits careful study by every thorough student of theoretical statistics.

FISHER, IRVING. The Best Form of Index Number. *Q. P. A. S. A.*, vol. 17 (1921).

—— *The Making of Index Numbers.* Boston, Houghton Mifflin, 1922.

—— *The Purchasing Power of Money.* N. Y., Macmillan, 1911.

FISHER, R. A. Frequency Distribution of the Values of Correlation Coefficients in Samples from an Indefinitely Large Population. *Biom.*, vol. 10 (1915).

—— On the Interpretation of χ^2 from Contingency Tables and the Calculation of P. *J. R. S. S.*, vol. 85 (1922).

—— On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. *Metron*, vol. 1, no. 4 (1921).

FLUX, A. W. Modes of Constructing Index Numbers. *Quar. Jour. Econ.*, vol. 21 (1907).

—— The Measurement of Price Changes. *J. R. S. S.*, vol. 84 (1921).

FORCHER, H. *Die statistiche Method als selbständige Wissenschaft.* Leipzig, Veit, 1913.

—— Zu den Anwendungen der Wahrscheinlichkeitsrechnung in der mathematischen Statistik. *Bull. Inst. Int. de Stat.*, vol. 20 (1916).

FOUNTAIN, H. The Construction of Index Numbers of Prices. Board of Trade *Report* on wholesale and retail prices in the United Kingdom, 1903.

GALTON, FRANCIS. Correlations and their Measurement. *Proc. Roy. Soc. Lond.*, vol. 45 (1888).

—— Grades and Deviates (including a table of normal deviates corresponding to each millesimal grade in the length of an array, and a figure). *Biom.*, vol. 5 (1906–07).

—— *Natural Inheritance.* London, Macmillan, 1889.

—— Regression towards Mediocrity in Hereditary Stature. *Jour. Anthrop. Inst.*, vol. 15 (1886).

—— Spurious Correlation due to Indices. *Proc. Roy. Soc.*, vol. 60 (1897).

—— Statistics by Intercomparison, with Remarks on the Law of Frequency of Error. *Phil. Mag.*, Series 4, vol. 44 (1875).

—— The Geometric Mean in Vital and Social Statistics. *Proc. Roy. Soc.*, vol. 29 (1879).

GEIGER, H. *Cf.* Rutherford, E.

GIBSON, WINIFRED. Tables for Facilitating the Computation of Probable Errors. *Biom.*, vol. 4 (1905–06).

GIFFEN, ROBERT. *Statistics, 1898–1900.* London, Macmillan, 1913.

GIUSTI, U. Sur la mesure de la densité des agglomérations urbaines en général et, en particulier, des agglomérations italiennes. *Bull. Inst. Int. de Stat.*, vol. 20 (1916).

GLAISHER, J. W. L. On the Law of Facility of Errors of Observations and on the Method of Least Squares. *Mem. R. A. S.*, vol. 36 (1872).

GLOVER, J. W. Derivation of the U. S. Mortality Table by Osculatory Interpolation. *Q. P. A. S. A.*, vol. 12 (1916).

—— *Tables of Applied Mathematics, Finance, Insurance, Statistics.* Ann Arbor, George Wahr (1923).

Contains an extensive set of tables, including probability functions.

—— *United States Life Tables*, 1890, 1901, 1910, and 1901–1910. Explanatory text, mathematical theory, computations, graphs, and original statistics. Washington, Government Printing Office, 1921.

GREENWOOD, M. On Errors of Random Sampling in Certain Cases not Suitable for the Application of a "Normal Curve of Frequency," *Biom.*, vol. 9 (1913).

GULDBERG, A. A Remark on Correlation. *Skand. Akt.* (1919).

—— Expansions Useful in the Theory of Frequency Distributions. *J. R. S. S.*, vol. 83 (1920).

—— On Correlation. *Norsk. Mat. Foreningssk.* Serie I, Nr. 5 (1921).

HARRIS, J. A. A Short Method of Calculating the Coefficient of Correlation in the Case of Integral Variates. *Biom.*, vol. 7 (1910).

—— On Inter-Class and Intra-Class Correlations. *Biom.*, vol. 9 (1913).

—— On the Spurious Values of the Intra-Class Correlation Coefficients Arising from Disorderly Differentiation within Classes. *Biom.*, vol. 10 (1915).

—— The Arithmetic of the Product Moment Method of Calculating the Coefficient of Correlation. *Amer. Naturalist*, vol. 44 (1910).

HENDERSON, ROBERT. A Practical Interpolation Formula with a Theoretical Introduction. *Trans. Actuarial Soc. Amer.*, vol. 9 (1905).

—— Graduation of Mortality and Other Tables. *Actuarial Studies*, No. 4 (1919).

HERON, D. *Cf.* Pearson, K.

HIGHAM, J. A. On the Adjustment or Graduation of Mortality Tables. *Jour. Inst. Actuaries*, vol. 23 (1882).

HOOKER, R. H. An Elementary Explanation of Correlation. *Q. Jour. Roy. Met. Soc.* (1918).

—— Correlation of Successive Observations. *J. R. S. S.*, vol. 68 (1905).

HUNTINGTON, E. V. Mathematics and Statistics, with an Elementary Account of the Correlation Coefficient and the Correlation Ratio. *Amer. Math. Monthly*, vol. 26 (1919).

—— *Handbook of Mathematics for Engineers.* N. Y., McGraw-Hill, 1918.

Contains tables of squares, cubes, and roots, and exponential functions likely to be useful to the statistician.

JACKSON, D. Note on the Median of a Set of Numbers. *Bull. Amer. Math. Soc.*, vol. 27 (1921).

JASTREMSKY. Der Auslese-koeffizient. *Zeits. für die gesamte Versich. Wiss.*, Bd. 12 (1912).

JEVONS, W. S. A Serious Fall in the Value of Gold Ascertained and its Social Effects Set Forth. (1863.) Reprinted in: *Investigations in Currency and Finance.* London, Macmillan, 1884.

—— *The Principles of Science.* London, Macmillan, 1907.

JOFFE, S. A. Interpolation Formulæ and Central-Difference Notation. *Trans. Actuarial Soc. Amer.*, vol. 18 (1917).

—— Parallel Proofs of Everett's, Gauss's, and Newton's Central Difference Interpolation Formulæ. *Trans. Actuarial Soc. Amer.*, vol. 20 (1919).

John, V. *Geschichte der Statistik.* Stuttgart, Enke, 1884.

Jones, D. C. *A First Course in Statistics.* London, G. Bell & Sons, 1921.

The book aims to be an introductory work, but assumes more extensive mathematical knowledge than is possessed by many beginners. A little more than half the book is devoted to a presentation of statistical fundamentals, up to and including simple correlation. Then follow valuable chapters on the elementary theory of sampling and the methods of curve fitting. The illustrations are mainly economic. To one who already has an introductory knowledge of statistics, the book should be very useful.

Kapteyn, J. C. Definition of the Correlation Coefficient. *Monthly Notices, R. A. S.*, vol. 72 (1912).

—— *Skew Frequency Curves.* Groningen, Noordhoff, 1904.

Karup, J. On a New Mechanical Method of Graduation. *Trans. Sec. Int. Cong. Actuaries* (1898).

Kelley, T. L. Certain Properties of Index Numbers. *Q. P. A. S. A.*, vol. 17 (1921).

—— *Chart to Facilitate the Calculation of Partial Coefficients of Correlation and Regression Coefficients.* Monograph, Stanford Univ., 1921.

—— *Statistical Method.* New York, Macmillan, 1923.

A treatise dealing with many phases of statistical analysis. The book is adapted both to the needs of students and practicing statisticians. While much of the book can be studied with profit by those with but little mathematical training, other parts require considerable preliminary mathematical knowledge.

Keynes, J. M. *A Treatise on Probability.* London, Macmillan, 1921.

A philosophic inquiry into the bases of the theory of probability. A critical discussion of fundamental principles rather than an exposition of method. While it offers little for the reader who desires merely to understand how to analyze statistical material, it will surely prove stimulating to him whose chief wish is to gain a clearer insight into the nature of statistical knowledge.

—— Principal Averages and the Laws of Error which Lead to Them. *J. R. S. S.*, vol. 74 (1911).

King, George. Notes on Summation Formulas of Graduation. *Jour. Inst. Actuaries*, vol. 41 (1907).

—— On a New Method of Constructing and Graduating Mortality and Other Tables. *Jour. Inst. Actuaries*, vol. 43 (1909).

—— On the Construction of Mortality Tables from Census Returns and Records of Deaths. *Jour. Inst. Actuaries*, vol. 42 (1908).

—— On the Error Introduced into Mortality Tables by Summation Formulas of Graduation. *Jour. Inst. Actuaries*, vol. 41 (1907).

King, W. I. *Elements of Statistical Method.* N. Y., Macmillan, 1918.

Knibbs, G. H. Prices, Price Indexes, and Cost of Living in Australia. Commonwealth Bureau of Census and Statistics, Labor and Industrial Branch, *Report* No. 1 (1912).

—— Price Indexes, Their Nature and Limitations, the Technique of Computing Them, and Their Application in Ascertaining the Purchasing Power of Money. Commonwealth Bureau of Census and Statistics, Labor and Industrial Branch, *Report* No. 9 (1918).

Koren, J. *History of Statistics.* N. Y., Macmillan, 1918.

Kries, J. von. *Die Principien der Wahrscheinlichkeitsrechnung. Eine logische Untersuchung.* Freiburg, 1886.

Landré, C. L. *Mathematisch-technische Kapitel zur Lebens-versicherung.* Jena, Fischer, 1901.

Laplace, P. S. *Théorie analytique des probabilités.* 3d ed., Paris, Courcier, 1820.

Lee, A. *Cf.* Pearson, K.

—— *Cf.* Soper, H. E.

Lexis, W. *Abhandlungen zur Theorie der Bevölkerungs- und Moralstatistik.* Jena, Fischer, 1903.

The book reprints certain of the fundamental papers on the Lexis theory of dispersion.

—— Über die Theorie der Stabilität statisticher Reihen. *Jahrb. für nat. ök. u. Stat.* (1) vol. 32 (1879).

—— Über die Wahrscheinlichkeitsrechnung und deren Anwendung auf die Statistik. *Jahrb. für nat. ök. u. Stat.* (2) vol. 13 (1886).

—— *Zur Theorie der Massenerscheinungen in der menschlichen Gesellschaft.* Freiburg, Wagner, 1877.

Lidstone, G. J. Alternative Demonstration of the Formula for Osculatory Interpolation. *Jour. Inst. Actuaries*, vol. 42 (1908).

—— On the Rationale of Formulas for Graduation by Summation. *Jour. Inst. Actuaries*, vol. 42 (1908).

Lipka, J. *Graphical and Mechanical Computation.* N. Y., Wiley, 1918.

Lipps, G. F. Die Bestimmung der Abhängigkeits zwischen den Merkmalen eines Gegenstandes. Leipzig, *Ber. d. math. phys. Kl. d. Kgl. sächs. Ges. d. Wiss.* (1905).

Loomis, Elias. *An Introduction to Practical Astronomy*, pp. 202–07.

Lubbock, J. W. On the Comparison of Various Tables of Annuities. *Trans. Camb. Phil. Soc.*, vol. 3. Reprinted in the *Jour. of the Institute of Actuaries*, vol. 5 (1855).

Macaulay, F. R. Making and Using of Index Numbers. *Amer. Econ. Rev.*, vol. 6 (1916).

March, Lucien. Les modes de mesure du mouvement général des prix. *Metron*, 1921.

Meitzen, A. History, Theory and Technique of Statistics. *Trans.*, by R. P. Falkner, Phila., *Amer. Acad. Pol. Soc. Sci.*, 1891.

Merriman, Mansfield. *Method of Least Squares.* N. Y., Wiley, 1910.

Miner, J. R. *Tables of* $\sqrt{1 - r^2}$ *and* $1 - r^2$ *for Use in Partial Correlation and in Trigonometry.* Baltimore, Johns Hopkins Press, 1922.

Mitchell, W. C. Index Numbers of Wholesale Prices in the United States and Foreign Countries. *U. S. Bureau of Labor Statistics, Bull.* 284, 1921. (Revision of *Bull.* 173, 1915.)

Moore, C. N. On the Coefficient of Correlation as a Measure of Relationship. *Science*, vol. 42 (1915).

Moore, H. L. *Economic Cycles; Their Law and Cause.* N. Y., Macmillan, 1914.

—— Generating Cycles. (3 articles.) *Quar. Jour. Econ.*, vol. 35 (1921).

Moore, L. B. *Cf.* Pearson, K.

Norton, J. P. *Statistical Studies in the New York Money Market.* N.Y., Macmillan, 1902.

Pearl, R. The Calculation of the Probable Errors of Certain Constants of the Normal Curve. *Biom.*, vol. 5 (1906–07).

Pearson, K. *Cf.* Blakeman, J.; Cave, B. M.; and Soper, H. E.

—— Contributions to the Mathematical Theory of Evolution. *Phil. Trans. Roy. Soc.*, vol. 185 A (1894).

—— "Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson"; A rejoinder. *Biom.*, vol. 4 (1905–06).

—— Note on the Significant or Non-Significant Character of a Sub-Sample drawn from a Sample. *Biom.*, vol., 5 (1906–07).

—— On a Correction Needful in the Case of the Correlation Ratio. *Biom.*, vol. 8 (1911).

—— On a Form of Spurious Correlation that may Arise when Indices are Used in the Measurement of Organs. *Proc. Roy. Soc.*, vol. 60 (1897).

—— On a New Method of Determining Correlation when One Variable is Given by Alternative and the Other by Multiple Categories. *Biom.*, vol. 7 (1910).

—— On a New Method of Determining Correlation between a Measured Character *A* and a Character *B* of which only the Percentage of Cases wherein *B* Exceeds (or Falls Short of) a Given Intensity is Recorded for Each Grade of *A*. *Biom.*, vol. 7 (1910).

—— On an Extension of the Method of Correlation by Grades or Ranks. *Biom.*, vol. 10 (1914).

—— On Certain Properties of the Hypergeometrical Series, and on the Fitting of Such Series to Observation Polygons in the Theory of Chance. *Phil. Mag.*, Series 5, vol. 47 (1899).

—— On Curves which are Most Suitable for Describing the Frequency of Random Samples of a Population. *Biom.*, vol. 5 (1905–06).

—— On Further Methods of Determining Correlation. *Drapers' Company Res. Mem., Biom. Series IV* (1907).

—— On Lines and Planes of Closest Fit to Systems of Points in Space. *Phil. Mag.*, Series 6, vol. 2 (1901).

—— On the Correction to be Made in the Correlation Ratio. *Biom.*, vol. 8 (1911).

—— On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling. *Phil. Mag.* (1900).

—— On the Dissection of Asymmetrical Frequency Curves. *Phil. Trans. Roy. Soc.*, vol. 185 A (1894).

—— On the Distribution of Standard Deviations of Small Samples. *Biom.*, vol. 10 (1915).

—— On the General Influence of Selection on Correlation and Variation. *Biom.*, vol. 8 (1912).

—— On the Mathematical Theory of Errors of Judgment, with Special Reference to the Personal Equation. *Phil. Trans. Roy. Soc.*, A, vol. 198 (1903).

—— On the Measurement of the Influence of "Broad Categories" on Correlation. *Biom.*, vol. 9 (1913).

—— On the Modal Value of an Organ or Character. *Biom.*, vol. 1 (1902).

—— On the Partial Correlation Ratio. *Proc. Roy. Soc. Lond.*, A, vol. 91 (1915).

—— On the Probable Error of the Biserial *r*. *Biom.*, vol. 11 (1916).

—— On the Probable Error of the Coefficient of Correlation as Found from a Fourfold Table. *Biom.*, vol. 9 (1913).

—— On the Probable Error of the Coefficient of Mean Square Contingency. *Biom.*, vol. 10 (1915).

Pearson, K. On the Systematic Fitting of Curves to Observations and Measurements. *Biom.*, vol. 1 (1901); vol. 2 (1902–03); vol. 4 (1905–06).

—— On the Theory of Contingency and its Relation to Association and Normal Correlation. *Drapers' Company Res. Mem., Biom. Series I* (1904).

—— On the Theory of Skew Correlation and Non-Linear Regression. *Drapers' Company Res. Mem., Biom. Series II* (1905).

—— On the χ^2 Test of Goodness of Fit. *Biom.*, vol. 14 (1922).

This paper is a rejoinder to Fisher and Yule.

—— Further Note on the χ^2 Test of Goodness of Fit. *Biom.*, vol. 14 (1923).

—— Professor von Torok's Attack on the Arithmetic Mean. *Biom.*, vol. 2 (1902–03).

—— Regression, Heredity and Panmixia. *Phil. Trans. Roy. Soc.*, A, vol. 187 (1896).

—— Second Note on the Coefficient of Correlation as Determined from the Quantitative Measurement of One Variate and the Ranking of a Second Variate. *Biom.*, vol. 13 (1920).

—— Skew Frequency Curves. A Rejoinder to Professor Kapteyn. *Biom.*, vol. 5 (1906–07).

—— Skew Variation in Homogeneous Material. *Phil. Trans. Roy. Soc.*, vol. 186 A (1895).

—— Second Supplement to a Memoir on Skew Variation. *Phil. Trans. Roy. Soc.*, vol. 216 A (1916).

—— Tables for Statisticians and Biometricians. Cambridge Univ. Press, 1914.

—— The Grammar of Science. 2d ed., London, A. & C. Black, 1900.

—— (Editorial.) On an Elementary Proof of Sheppard's Formulæ for Correcting Raw Moments and on Other Allied Points. *Biom.*, vol. 3 (1904).

—— (Editorial.) On the Probable Errors of the Frequency Constants. *Biom.*, vol. 2 (1902–03).

Pearson, K., and Filon, L. N. G. On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation. *Phil. Trans. Roy. Soc.*, A, vol. 191 (1898).

Pearson, K., and Heron, D. On the Theories of Association. *Biom.*, vol. 9 (1913).

Pearson, K., Moore, L. B., Filon, L. N. G., and Lee, A. On the Correlation of Characters not Quantitatively Measurable. *Phil. Trans. Roy. Soc.*, A, vol. 195 (1901).

Pearson, K., and Young, A. W. On the Probable Error of a Coefficient of Contingency without Approximations. *Biom.*, vol. 11 (1915).

Pearson, K., and Elderton, Ethel M. On the Variate Difference Method. *Biom.*, vol. 14 (1923).

Persons, W. M. Fisher's Formula for Index Numbers. *Rev. Econ. Stat.*, Prel. vol. 3 (1921).

—— *Indices of General Business Conditions.* Cambridge, Harvard Univ. Press, 1919.

—— (Transl.) Statistical Averages. Trans. from Franz Zizek's *Die statistischen Mittelwerthe*, with additional notes and references. N. Y., Henry Holt, 1913.

A non-mathematical discussion of the properties and the limitations of the several measures of average tendency and of dispersion. The detailed critical remarks on the different types of averages and measures of dispersion render the book particularly valuable to the beginner in statistical analysis.

PERSONS, W. M. The Variate Difference Method and Curve-Fitting. *Q. P. A. S. A.*, vol. 15 (1917).

PIGOU, A. C. *The Economics of Welfare.* London, Macmillan, 1920.

POISSON, S. D. *Recherches sur la probabilité des jugements.* Paris, Bachelier, 1837.

—— Sur la proportion des naissances des filles et des garçons. *Mem. de l'Acad. des Sci.*, vol. 9 (1829).

QUETELET, A. *Lettres sur la théorie des probabilités. Trans.* by O. G. Downes. London, C. & E. Layton, 1849.

RICE, H. L. *The Theory and Practice of Interpolations.* Lynn, Mass., 1899.

The book contains the derivations of the various interpolation formulas and their numerical applications to problems in practical astronomy.

RIETZ, H. L. On Functional Relations for which the Coefficient of Correlation is Zero. *Q. P. A. S. A.*, vol. 16 (1919).

—— On the Mathematical Theory of Risk and Landre's Theory of the Maximum. *The Record, Amer. Inst. Actuaries*, vol. II (1913).

—— On the Theory of Correlation with Special Reference to Certain Significant Loci on the Plane of Distribution in the Case of Normal Correlation. *Annals of Math.*, vol. 13 (1912).

—— Statistical Methods. Appendix to E. Davenport's *Principles of Breeding.* N. Y., Ginn, 1907.

—— Urn Schemata as a Basis for the Development of Correlation Theory. *Annals of Math.*, vol. 21 (1920).

RIETZ, H. L., and SHADE, IMOGENE. Correlation of Efficiency in Mathematics and Efficiency in Other Subjects. *U. of Ill. Studies*, vol. 6 (1908).

RITCHIE-SCOTT, A. The Correlation Coefficient of a Polychoric Table. *Biom.*, vol. 12 (1918).

RUGG, H. O. *Statistical Methods Applied to Education.* Boston, Houghton Mifflin, 1917.

RUNNING, T. R. *Empirical Formulas.* N. Y., Wiley, 1917.

RUTHERFORD, E., and GEIGER, H. The Probability Variations in the Distribution of α Particles. *Phil. Mag.*, Series 6, vol. 20 (1910).

SCHUSTER, A. Periodicities of Sunspots. *Phil. Trans. Roy. Soc.*, A, vol. 200 (1906).

—— The Investigation of Hidden Periodicities. *Terr. Magn.*, vol. 3 (1898).

—— The Periodogram and its Optical Analogy. *Proc. Roy. Soc.*, A, vol. 77 (1906).

SHADE, I. *Cf.* Rietz, H. L.

SHEARER, G. *Cf.* Carse, G. A.

SECRIST, H. *Introduction to Statistical Methods.* N. Y., Macmillan, 1917.

SHEPPARD, W. F. Central Difference Formulæ. *Proc. Lond. Math. Soc.*, vol. 31 (1899).

—— Central Difference Interpolation Formulæ. *Jour. Inst. Actuaries*, vol. 50 (1916).

—— New Tables of the Probability Integral. *Biom.*, vol. 2 (1902).

—— On the Calculation of the Probable Values of the Frequency Constants, from Data Arranged according to Equidistant Divisions of a Scale. *Proc. Lond. Math. Soc.*, vol. 29 (1898).

Sheppard, W. F. On the Application of the Theory of Error to Cases of Normal Distribution and Normal Correlation. *Phil. Trans. Roy. Soc.*, A, vol. 192 (1898).

—— The Calculation of the Moments of a Frequency Distribution. *Biom.*, vol. 5 (1906–07).

Slutsky, E. On the Criterion of Goodness of Fit of Regression Lines and on the Best Method of Fitting them to Data. *J. R. S. S.*, vol. 77 (1913).

Soper, H. E. On the Probable Error of the Biserial Expression for the Correlation Coefficient. *Biom.*, vol. 10 (1915).

—— On the Probable Error of the Correlation Coefficient to a Second Approximation. *Biom.*, vol. 9 (1913).

—— Tables of Poisson's Exponential Binomial Limit. *Biom.*, vol. 10 (1914).

Soper, H. E., Young, A. W., Cave, B. M., Lee, A., and Pearson, K. On the Distribution of the Correlation Coefficient in Small Samples. *Biom.*, vol. 11 (1916).

Spearman, C. Correlation Calculated from Faulty Data. *Brit. Jour. Psych.*, vol. 3 (1910).

—— Demonstration of Formulæ for Time Measurement of Correlation. *Amer. Jour. Psych.*, vol. 18 (1907).

—— Die praktische Elimination des Einflusses des zufälligen Fehler von dem Korrelationskoeffizienten. *Zeitsch. für angew. Psych.*, vol. 6 (1912).

—— Footrule for Measuring Correlation. *Brit. Jour. Psych.*, vol. 2 (1909).

—— The Proof and Measurement of Association between Two Things. *Amer. Jour. Psych.*, vol. 15 (1904).

Spencer, J. On the Graduation of the Rates of Sickness, etc. *Jour. Inst. Actuaries*, vol. 38 (1904).

Sprague, T. B. Explanation of a New Formula for Interpolation. *Jour. Inst. Actuaries*, vol. 22 (1880).

Student. An Experimental Determination of the Probable Error of Dr. Spearman's Correlation Coefficients. *Biom.*, vol. 13 (1920).

—— On the Error of Counting with a Hæmacytometer. *Biom.*, vol. 5 (1907).

—— On the Probable Error of a Correlation Coefficient. *Biom.*, vol. 6 (1908).

—— The Correction to be Made to the Correlation Ratio for Grouping. *Biom.*, vol. 9 (1913).

—— The Elimination of Spurious Correlation due to Position in Time and Space. *Biom.*, vol. 10 (1912).

—— The Probable Error of the Mean. *Biom.*, vol. 6 (1908).

Thiele, T. N. *Theory of Observations.* London, C. & E. Layton, 1903.

Thompson, A. J. Table of Coefficients of Everett's Central Difference Interpolation Formula. *Tracts for Computers*, ed. Karl Pearson, No. 5 (1921).

Thomson, G. H. *Cf.* Brown, W.

Thorndike, E. L. *An Introduction to the Theory of Mental and Social Measurement.* N. Y., Science Press, 1904.

Todhunter, I. *History of the Mathematical Theory of Probability, from the Time of Pascal to that of Laplace.* Cambridge, 1865.

Trachtenberg, M. I. A Note on a Property of the Median. *J. R. S. S.*, vol. 78 (1915).

Tschuprow, A. A. Die Aufgaben der Theorie der Statistik. *Jahrb. für gesetzg. Verwalt. u. Volkswirtsch.*, vol. 29 (1905).

U. S. Bureau of Labor Statistics. *Cf.* Mitchell, W. C.

U. S. Bureau of the Census. *Statistical Atlas of the United States*, 1900.

VENN, J. On the Nature and Uses of Averages. *J. R. S. S.*, vol. 54 (1891).

—— *The Logic of Chance.* 3d ed., London, Macmillan, 1888.

VIGOR, H. D., and YULE, G. U. On the Sex-Ratios of Births in the Registration Districts of England and Wales, 1881–1890. *J. R. S. S.*, vol. 69 (1906).

WALLACE, H. A. *Agricultural Prices.* Des Moines, 1920.

WALSH, C. M. *The Measurement of General Exchange-Value.* N. Y., Macmillan, 1901.

—— *The Problem of Estimation.* London, King, 1921.

WELD, L. D. *Theory of Errors and Least Squares.* N. Y., Macmillan, 1916.

WEST, C. J. *Introduction to Mathematical Statistics.* Columbus, R. G. Adams, 1918.

WESTERGAARD, H. *Die Grundzüge der Theorie der Statistik.* Jena, Fischer, 1890.

WHIPPLE, G. C. *Vital Statistics.* N. Y., Wiley, 1922.

WHITTAKER, LUCY. On Poisson's Law of Small Numbers. *Biom.*, vol. 10 (1914).

WICKSELL, S. D. An Exact Formula for Spurious Correlation. *Metron*, vol. 1 (1920).

WISSLER, C. The Spearman Correlation Formula. *Science*, N. S., vol. 22 (1905).

WOOLHOUSE, W. S. B. On Interpolation, Summation, and the Adjustment of Numerical Tables. *Jour. Inst. Actuaries*, vols. 11, 12 (1863–1865).

—— Explanation of a New Method of Adjusting Mortality Tables. *Jour. Inst. Actuaries*, vol. 15 (1870).

—— On an Improved Theory of Annuities and Assurances. *Jour. Inst. Actuaries*, vol. 15 (1870).

WRIGHT, P. G. Moore's Work in Cycles. *Quar. Jour. Econ.*, vol. 36 (1922).

YOUNG, A. A. Fisher's The Making of Index Numbers. *Quar. Jour. Econ.*, vol. 37 (1923).

—— The Measurement of Changes of the General Price Level. *Quar. Jour. Econ.*, vol. 35 (1921).

YOUNG, A. W. *Cf.* Pearson, K.

YULE, G. U. *Cf.* Vigor, H. D.

YULE, G. U. *An Introduction to the Theory of Statistics.* 5th ed., London, Griffin, 1919.

A presentation of many of the chief methods of statistics. The mathematical work is planned to avoid the calculus, and the illustrations are chiefly from biology. Almost indispensable for the student and practicing statistician.

—— Fluctuations of Sampling in Mendelian Ratios. *Proc. Camb. Phil. Soc.*, vol. 17 (1904).

—— Notes on the History of Pauperism in England and Wales, etc. (Supplementary notes on the determination of the mode.) *J. R. S. S.*, vol. 59 (1896).

—— Notes on the Theory of Association of Attributes in Statistics. *Biom.*, vol. 2 (1903).

—— On the Application of the χ^2 Method to Association and Contingency Tables, with Experimental Illustrations. *J. R. S. S.*, vol. 85 (1922).

—— On the Association of Attributes in Statistics. *Phil. Trans. Roy. Soc.*, A, vol. 194 (1900).

—— On the Methods of Measuring Association between Two Attributes. *J. R. S. S.*, vol. 75 (1912).

Yule, G. U. On the Interpretation of Correlation between Indices and Ratios. *J. R. S. S.*, vol. 73 (1910).

—— On the Theory of Correlation. *J. R. S. S.*, vol. 60 (1897).

—— On the Theory of Correlation for any Number of Variables Treated by a New System of Notation. *Proc. Roy. Soc. Lond.*, vol. 79 A (1907).

—— On the Time Correlation Problem, with Especial Reference to the Variate Difference Correlation Method. *J. R. S. S.*, vol. 84 (1921).

—— The Applications of the Method of Correlation to Social and Economic Statistics. *J. R. S. S.*, vol. 72 (1909).

—— The Significance of Bravais' Formulæ for Regression in the Case of Skew Variation. *Proc. Roy. Soc. Lond.*, vol. 60 (1897).

Zizek, F. *Grundriss der Statistik.* Leipzig, Duncker und Humblot, 1921.

TABLES OF PROBABILITY FUNCTIONS

Areas under the Curve $y = \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, the Function $\phi(t)$, and its Second, Third, and Fourth Derivatives[1]

t	$\int_0^t \phi(t)dt$	$\phi(t)$	$\phi^{(2)}(t)$	$\phi^{(3)}(t)$	$\phi^{(4)}(t)$	t
.00	.000000	.39894	− .39894	.00000	1.19683	.00
.01	.003989	.39892	− .39888	.01197	1.19653	.01
.02	.007978	.39886	− .39870	.02393	1.19563	.02
.03	.011966	.39876	− .39840	.03588	1.19414	.03
.04	.015953	.39862	− .39799	.04781	1.19204	.04
.05	.019938	.39844	− .39745	.05972	1.18936	.05
.06	.023922	.39822	− .39679	.07159	1.18608	.06
.07	.027903	.39797	− .39602	.08344	1.18221	.07
.08	.031881	.39767	− .39512	.09524	1.17775	.08
.09	.035856	.39733	− .39411	.10699	1.17271	.09
.10	.039828	.39695	− .39298	.11869	1.16708	.10
.11	.043795	.39654	− .39174	.13033	1.16088	.11
.12	.047758	.39608	− .39038	.14190	1.15410	.12
.13	.051717	.39559	− .38890	.15341	1.14676	.13
.14	.055670	.39505	− .38731	.16484	1.13885	.14
.15	.059618	.39448	− .38560	.17618	1.13038	.15
.16	.063560	.39387	− .38379	.18744	1.12137	.16
.17	.067495	.39322	− .38186	.19861	1.11180	.17
.18	.071424	.39253	− .37981	.20968	1.10170	.18
.19	.075345	.39181	− .37766	.22064	1.09106	.19
.20	.079260	.39104	− .37540	.23150	1.07990	.20
.21	.083166	.39024	− .37303	.24224	1.06823	.21
.22	.087064	.38940	− .37056	.25286	1.05604	.22
.23	.090954	.38853	− .36798	.26336	1.04335	.23
.24	.094835	.38762	− .36529	.27373	1.03018	.24
.25	.098706	.38667	− .36250	.28396	1.01651	.25
.26	.102568	.38568	− .35961	.29405	1.00238	.26
.27	.106420	.38466	− .35662	.30401	0.98778	.27
.28	.110261	.38361	− .35353	.31381	0.97273	.28
.29	.114092	.38251	− .35035	.32346	0.95723	.29
.30	.117911	.38139	− .34706	.33295	0.94130	.30
.31	.121720	.38023	− .34369	.34228	0.92495	.31
.32	.125516	.37903	− .34022	.35145	0.90819	.32
.33	.129300	.37780	− .33666	.36045	0.89103	.33
.34	.133072	.37654	− .33301	.36927	0.87348	.34
.35	.136831	.37524	− .32927	.37791	0.85555	.35
.36	.140576	.37391	− .32545	.38638	0.83726	.36
.37	.144309	.37255	− .32155	.39466	0.81862	.37
.38	.148027	.37115	− .31756	.40275	0.79963	.38
.39	.151732	.36973	− .31349	.41065	0.78032	.39
.40	.155422	.36827	− .30935	.41835	0.76070	.40
.41	.159097	.36678	− .30586	.42586	0.74077	.41
.42	.162757	.36526	− .30083	.43317	0.72056	.42
.43	.166402	.36371	− .29646	.44027	0.70007	.43
.44	.170031	.36213	− .29203	.44717	0.67932	.44
.45	.173645	.36053	− .28752	.45386	0.65832	.45
.46	.177242	.35889	− .28295	.46034	0.63709	.46
.47	.180822	.35723	− .27831	.46660	0.61564	.47
.48	.184386	.35553	− .27362	.47265	0.59398	.48
.49	.187933	.35381	− .26886	.47848	0.57213	.49

[1] The values of $\phi(t)$ and of its derivatives are taken from *Tables of Applied Mathematics*, by James W. Glover, through the courtesy of the publisher, George Wahr.

Areas under the Curve $y = \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, the Function $\phi(t)$, and its Second, Third, and Fourth Derivatives

t	$\int_0^t \phi(t)dt$	$\phi(t)$	$\phi^{(2)}(t)$	$\phi^{(3)}(t)$	$\phi^{(4)}(t)$	t
.50	.191462	.35207	− .26405	.48409	.55010	.50
.51	.194974	.35029	− .25918	.48948	.52791	.51
52	.198468	.34849	− .25426	.49465	.50556	.52
.53	.201944	.34667	− .24929	.49959	.48308	.53
54	.205402	.34482	− .24427	.50431	.46048	.54
.55	.208840	.34294	− .23920	.50880	.43777	.55
.56	.212260	.34105	− .23409	.51306	.41497	.56
.57	.215661	.33912	− .22894	.51710	.39208	.57
.58	.219043	.33718	− .22375	.52091	.36913	.58
.59	.222405	.33521	− .21853	.52448	.34613	.59
.60	.225747	.33322	− .21326	.52783	.32309	.60
.61	.229069	.33121	− .20797	.53094	.30003	.61
.62	.232371	.32918	− .20265	.53383	.27696	.62
.63	.235653	.32713	− .19729	.53648	.25390	.63
.64	.238914	.32506	− .19192	.53891	.23085	.64
.65	.242154	.32297	− .18652	.54110	.20783	.65
.66	.245373	.32086	− .18110	.54306	.18486	.66
.67	.248571	.31874	− .17566	.54480	.16195	.67
.68	.251748	.31659	− .17020	.54630	.13912	.68
.69	.254903	.31443	− .16473	.54758	.11636	.69
.70	.258036	.31225	− .15925	.54863	.09371	.70
.71	.261148	.31006	− .15376	.54945	.07116	.71
.72	.264238	.30785	− .14826	.55005	.04874	.72
.73	.267305	.30563	− .14276	.55043	.02646	.73
.74	.270350	.30339	− .13725	.55058	.00433	.74
.75	.273373	.30114	− .13175	.55052	− .01764	.75
.76	.276373	.29887	− .12624	.55023	− .03944	.76
.77	.279350	.29659	− .12074	.54973	− .06106	.77
.78	.282305	.29431	− .11525	.54901	− .08248	.78
.79	.285236	.29200	− .10976	.54808	− .10369	.79
.80	.288145	.28969	− .10429	.54694	− .12468	.80
.81	.291030	.28737	− .09883	.54559	− .14545	.81
.82	.293892	.28504	− .09338	.54403	− .16597	.82
.83	.296731	.28269	− .08795	.54227	− .18624	.83
.84	.299546	.28034	− .08253	.54031	− .20626	.84
.85	.302338	.27798	− .07714	.53814	− .22600	.85
.86	.305106	.27562	− .07177	.53579	− .24546	.86
.87	.307850	.27324	− .06643	.53324	− .26464	.87
.88	.310570	.27086	− .06111	.53049	− .28351	.88
.89	.313267	.26848	− .05582	.52757	− .30208	.89
.90	.315940	.26609	− .05056	.52445	− .32034	.90
.91	.318589	.26369	− .04533	.52116	− .33827	.91
.92	.321214	.26129	− .04013	.51769	− .35587	.92
.93	.323814	.25888	− .03497	.51404	− .37314	.93
.94	.326391	.25647	− .02985	.51023	− .39005	.94
.95	.328944	.25406	− .02477	.50624	− .40662	.95
.96	.331472	.25164	− .01973	.50210	− .42283	.96
.97	.333977	.24923	− .01473	.49779	− .43867	.97
.98	.336457	.24681	− .00977	.49332	− .45414	.98
.99	.338913	.24439	− .00486	.48871	− .46923	.99

Areas under the Curve $y = \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, the Function $\phi(t)$, and its Second, Third, and Fourth Derivatives

t	$\int_0^t \phi(t)dt$	$\phi(t)$	$\phi^{(2)}(t)$	$\phi^{(3)}(t)$	$\phi^{(4)}(t)$	t
1.00	.341345	.24197	.00000	.48394	− .48394	1.00
1.01	.343752	.23955	.00482	.47903	− .49827	1.01
1.02	.346136	.23713	.00958	.47398	− .51220	1.02
1.03	.348495	.23471	.01429	.46879	− .52573	1.03
1.04	.350830	.23230	.01896	.46346	− .53887	1.04
1.05	.353141	.22988	.02356	.45801	− .55160	1.05
1.06	.355428	.22747	.02812	.45243	− .56393	1.06
1.07	.357690	.22506	.03261	.44673	− .57584	1.07
1.08	.359929	.22265	.03705	.44092	− .58734	1.08
1.09	.362143	.22025	.04143	.43499	− .59843	1.09
1.10	.364334	.21785	.04575	.42895	− .60909	1.10
1.11	.366500	.21546	.05001	.42281	− .61934	1.11
1.12	.368643	.21307	.05420	.41657	− .62917	1.12
1.13	.370762	.21069	.05834	.41023	− .63857	1.13
1.14	.372857	.20831	.06241	.40380	− .64755	1.14
1.15	.374928	.20594	.06641	.39728	− .65611	1.15
1.16	.376976	.20357	.07035	.39067	− .66425	1.16
1.17	.379000	.20121	.07423	.38399	− .67196	1.17
1.18	.381000	.19886	.07803	.37724	− .67924	1.18
1.19	.382977	.19652	.08177	.37041	− .68610	1.19
1.20	.384930	.19419	.08544	.36352	− .69255	1.20
1.21	.386861	.19186	.08904	.35656	− .69857	1.21
1.22	.388768	.18954	.09257	.34955	− .70417	1.22
1.23	.390651	.18724	.09603	.34248	− .70935	1.23
1.24	.392512	.18494	.09942	.33536	− .71411	1.24
1.25	.394350	.18265	.10274	.32820	− .71847	1.25
1.26	.396165	.18037	.10599	.32099	− .72241	1.26
1.27	.397958	.17810	.10916	.31375	− .72594	1.27
1.28	.399727	.17585	.11226	.30648	− .72907	1.28
1.29	.401475	.17360	.11529	.29917	− .73180	1.29
1.30	.403200	.17137	.11824	.29184	− .73413	1.30
1.31	.404902	.16915	.12113	.28449	− .73606	1.31
1.32	.406582	.16694	.12393	.27712	− .73760	1.32
1.33	.408241	.16474	.12667	.26974	− .73876	1.33
1.34	.409877	.16256	.12933	.26235	− .73953	1.34
1.35	.411492	.16038	.13192	.25495	− .73993	1.35
1.36	.413085	.15822	.13443	.24755	− .73995	1.36
1.37	.414656	.15608	.13687	.24015	− .73961	1.37
1.38	.416207	.15395	.13923	.23276	− .73890	1.38
1.39	.417736	.15183	.14152	.22537	− .73784	1.39
1.40	.419243	.14973	.14374	.21800	− .73642	1.40
1.41	.420730	.14764	.14588	.21065	− .73466	1.41
1.42	.422196	.14556	.14795	.20331	− .73256	1.42
1.43	.423642	.14350	.14995	.19600	− .73012	1.43
1.44	.425066	.14146	.15187	.18871	− .72736	1.44
1.45	.426471	.13943	.15372	.18145	− .72427	1.45
1.46	.427855	.13742	.15550	.17423	− .72087	1.46
1.47	.429219	.13542	.15721	.16704	− .71716	1.47
1.48	.430563	.13344	.15884	.15988	− .71315	1.48
1.49	.431889	.13147	.16040	.15277	− .70885	1.49

Areas under the Curve $y = \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, the Function $\phi(t)$, and its Second, Third, and Fourth Derivatives

t	$\int_0^t \phi(t)dt$	$\phi(t)$	$\phi^{(2)}(t)$	$\phi^{(3)}(t)$	$\phi^{(4)}(t)$	t
1.50	.433193	.12952	.16190	.14571	−.70425	1.50
1.51	.434478	.12758	.16332	.13869	−.69938	1.51
1.52	.435744	.12566	.16467	.13172	−.69423	1.52
1.53	.436992	.12376	.16595	.12481	−.68881	1.53
1.54	.438220	.12188	.16717	.11795	−.68314	1.54
1.55	.439429	.12001	.16831	.11114	−.67721	1.55
1.56	.440620	.11816	.16939	.10440	−.67104	1.56
1.57	.441792	.11632	.17040	.09772	−.66463	1.57
1.58	.442947	.11450	.17135	.09111	−.65799	1.58
1.59	.444083	.11270	.17222	.08456	−.65113	1.59
1.60	.445201	.11092	.17304	.07809	−.64405	1.60
1.61	.446301	.10915	.17379	.07168	−.63677	1.61
1.62	.447384	.10741	.17447	.06535	−.62928	1.62
1.63	.448449	.10567	.17509	.05910	−.62161	1.63
1.64	.449497	.10396	.17565	.05292	−.61375	1.64
1.65	.450528	.10226	.17615	.04682	−.60571	1.65
1.66	.451543	.10059	.17659	.04081	−.59751	1.66
1.67	.452540	.09893	.17697	.03487	−.58914	1.67
1.68	.453521	.09728	.17729	.02903	−.58063	1.68
1.69	.454486	.09566	.17755	.02326	−.57196	1.69
1.70	.455434	.09405	.17775	.01759	−.56316	1.70
1.71	.456367	.09246	.17790	.01200	−.55422	1.71
1.72	.457284	.09089	.17799	.00650	−.54516	1.72
1.73	.458185	.08933	.17803	.00110	−.53599	1.73
1.74	.459070	.08780	.17802	−.00422	−.52671	1.74
1.75	.459941	.08628	.17795	−.00944	−.51733	1.75
1.76	.460796	.08478	.17783	−.01456	−.50785	1.76
1.77	.461636	.08329	.17766	−.01959	−.49829	1.77
1.78	.462462	.08183	.17744	−.02453	−.48865	1.78
1.79	.463273	.08038	.17717	−.02937	−.47893	1.79
1.80	.464070	.07895	.17685	−.03411	−.46915	1.80
1.81	.464852	.07754	.17648	−.03875	−.45932	1.81
1.82	.465620	.07614	.17607	−.04329	−.44943	1.82
1.83	.466375	.07477	.17562	−.04774	−.43950	1.83
1.84	.467116	.07341	.17512	−.05208	−.42953	1.84
1.85	.467843	.07206	.17458	−.05633	−.41953	1.85
1.86	.468557	.07074	.17399	−.06047	−.40950	1.86
1.87	.469258	.06943	.17337	−.06452	−.39946	1.87
1.88	.469946	.06814	.17270	−.06846	−.38940	1.88
1.89	.470621	.06687	.17200	−.07231	−.37934	1.89
1.90	.471283	.06562	.17126	−.07605	−.36928	1.90
1.91	.471933	.06438	.17048	−.07969	−.35923	1.91
1.92	.472571	.06316	.16966	−.08323	−.34918	1.92
1.93	.473197	.06195	.16881	−.08667	−.33916	1.93
1.94	.473810	.06077	.16793	−.09002	−.32916	1.94
1.95	.474412	.05959	.16701	−.09326	−.31919	1.95
1.96	.475002	.05844	.16607	−.09640	−.30925	1.96
1.97	.475581	.05730	.16509	−.09944	−.29936	1.97
1.98	.476148	.05618	.16408	−.10239	−.28950	1.98
1.99	.476704	.05508	.16304	−.10523	−.27970	1.99

Areas under the Curve $y = \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, the Function $\phi(t)$, and its Second, Third, and Fourth Derivatives

t	$\int_0^t \phi(t)dt$	$\phi(t)$	$\phi^{(2)}(t)$	$\phi^{(3)}(t)$	$\phi^{(4)}(t)$	t
2.00	.477250	.05399	.16197	− .10798	− .26996	2.00
2.01	.477784	.05292	.16088	− .11063	− .26027	2.01
2.02	.478308	.05186	.15976	− .11319	− .25064	2.02
2.03	.478822	.05082	.15862	− .11565	− .24109	2.03
2.04	.479325	.04980	.15745	− .11801	− .23160	2.04
2.05	.479818	.04879	.15626	− .12028	− .22220	2.05
2.06	.480301	.04780	.15504	− .12245	− .21287	2.06
2.07	.480774	.04682	.15381	− .12454	− .20363	2.07
2.08	.481237	.04586	.15255	− .12653	− .19448	2.08
2.09	.481691	.04491	.15128	− .12843	− .18542	2.09
2.10	.482136	.04398	.14998	− .13024	− .17646	2.10
2.11	.482571	.04307	.14867	− .13196	− .16759	2.11
2.12	.482997	.04217	.14735	− .13359	− .15883	2.12
2.13	.483414	.04128	.14600	− .13513	− .15017	2.13
2.14	.483823	.04041	.14464	− .13659	− .14162	2.14
2.15	.484222	.03955	.14327	− .13797	− .13318	2.15
2.16	.484614	.03871	.14188	− .13926	− .12486	2.16
2.17	.484997	.03788	.14049	− .14046	− .11665	2.17
2.18	.485371	.03706	.13907	− .14159	− .10856	2.18
2.19	.485738	.03626	.13765	− .14263	− .10059	2.19
2.20	.486097	.03547	.13622	− .14360	− .09274	2.20
2.21	.486447	.03470	.13478	− .14449	− .08502	2.21
2.22	.486791	.03394	.13333	− .14530	− .07743	2.22
2.23	.487126	.03319	.13188	− .14604	− .06996	2.23
2.24	.487454	.03246	.13041	− .14670	− .06263	2.24
2.25	.487776	.03174	.12894	− .14729	− .05542	2.25
2.26	.488089	.03103	.12747	− .14781	− .04835	2.26
2.27	.488396	.03034	.12599	− .14826	− .04141	2.27
2.28	.488696	.02965	.12450	− .14864	− .03461	2.28
2.29	.488989	.02898	.12301	− .14895	− .02794	2.29
2.30	.489276	.02333	.12152	− .14920	− .02141	2.30
2.31	.489556	.02768	.12003	− .14938	− .01502	2.31
2.32	.489830	.02705	.11854	− .14950	− .00877	2.32
2.33	.490097	.02643	.11704	− .14956	− .00265	2.33
2.34	.490358	.02582	.11554	− .14955	.00332	2.34
2.35	.490613	.02522	.11405	− .14949	.00915	2.35
2.36	.490862	.02463	.11256	− .14937	.01485	2.36
2.37	.491106	.02406	.11106	− .14919	.02040	2.37
2.38	.491344	.02349	.10957	− .14896	.02582	2.38
2.39	.491576	.02294	.10808	− .14868	.03109	2.39
2.40	.491802	.02239	.10660	− .14834	.03623	2.40
2.41	.492024	.02186	.10512	− .14795	.04122	2.41
2.42	.492240	.02134	.10364	− .14752	.04608	2.42
2.43	.492451	.02083	.10217	− .14703	.05079	2.43
2.44	.492656	.02033	.10070	− .14650	.05537	2.44
2.45	.492857	.01984	.09924	− .14593	.05981	2.45
2.46	.493053	.01936	.09778	− .14531	.06411	2.46
2.47	.493244	.01889	.09633	− .14464	.06828	2.47
2.48	.493431	.01842	.09489	− .14394	.07231	2.48
2.49	.493613	.01797	.09345	− .14320	.07621	2.49

Areas under the Curve $y = \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, the Function $\phi(t)$, and its Second, Third, and Fourth Derivatives

t	$\int_0^t \phi(t)dt$	$\phi(t)$	$\phi^{(2)}(t)$	$\phi^{(3)}(t)$	$\phi^{(4)}(t)$	t
2.50	.493790	.01753	.09202	− .14242	.07997	2.50
2.51	.493963	.01709	.09060	− .14160	.08360	2.51
2.52	.494132	.01667	.08919	− .14075	.08710	2.52
2.53	.494297	.01625	.08779	− .13986	.09047	2.53
2.54	.494457	.01585	.08639	− .13894	.09372	2.54
2.55	.494614	.01545	.08501	− .13798	.09683	2.55
2.56	.494766	.01506	.08364	− .13700	.09982	2.56
2.57	.494915	.01468	.08227	− .13599	.10268	2.57
2.58	.495060	.01431	.08092	− .13495	.10542	2.58
2.59	.495201	.01394	.07957	− .13388	.10804	2.59
2.60	.495339	.01358	.07824	− .13279	.11053	2.60
2.61	.495473	.01323	.07692	− .13167	.11291	2.61
2.62	.495604	.01289	.07560	− .13053	.11517	2.62
2.63	.495731	.01256	.07431	− .12937	.11732	2.63
2.64	.495855	.01223	.07302	− .12818	.11935	2.64
2.65	.495975	.01191	.07174	− .12698	.12127	2.65
2.66	.496093	.01160	.07048	− .12576	.12308	2.66
2.67	.496207	.01130	.06923	− .12452	.12479	2.67
2.68	.496319	.01100	.06799	− .12326	.12638	2.68
2.69	.496427	.01071	.06676	− .12199	.12787	2.69
2.70	.496533	.01042	.06555	− .12071	.12926	2.70
2.71	.496636	.01014	.06435	− .11941	.13055	2.71
2.72	.496736	.00987	.06316	− .11810	.13174	2.72
2.73	.496833	.00961	.06199	− .11677	.13283	2.73
2.74	.496928	.00935	.06082	− .11544	.13383	2.74
2.75	.497020	.00909	.05968	− .11410	.13473	2.75
2.76	.497110	.00885	.05854	− .11274	.13555	2.76
2.77	.497197	.00861	.05742	− .11139	.13627	2.77
2.78	.497282	.00837	.05631	− .11002	.13691	2.78
2.79	.497365	.00814	.05522	− .10865	.13746	2.79
2.80	.497445	.00792	.05414	− .10727	.13793	2.80
2.81	.497523	.00770	.05308	− .10589	.13832	2.81
2.82	.497599	.00748	.05202	− .10450	.13863	2.82
2.83	.497673	.00727	.05099	− .10312	.13886	2.83
2.84	.497744	.00707	.04996	− .10173	.13902	2.84
2.85	.497814	.00687	.04895	− .10034	.13910	2.85
2.86	.497882	.00668	.04795	− .09895	.13912	2.86
2.87	.497948	.00649	.04697	− .09755	.13906	2.87
2.88	.498012	.00631	.04600	− .09616	.13894	2.88
2.89	.498074	.00613	.04505	− .09478	.13875	2.89
2.90	.498134	.00595	.04411	− .09339	.13850	2.90
2.91	.498193	.00578	.04318	− .09201	.13819	2.91
2.92	.498250	.00562	.04227	− .09063	.13782	2.92
2.93	.498305	.00545	.04137	− .08925	.13739	2.93
2.94	.498359	.00530	.04048	− .08788	.13691	2.94
2.95	.498411	.00514	.03961	− .08651	.13638	2.95
2.96	.498462	.00499	.03875	− .08515	.13579	2.96
2.97	.498511	.00485	.03791	− .08380	.13515	2.97
2.98	.498559	.00471	.03708	− .08245	.13346	2.98
2.99	.498605	.00457	.03626	− .08111	.13373	2.99

Areas under the Curve $y = \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, the Function $\phi(t)$, and its Second, Third, and Fourth Derivatives

t	$\int_0^t \phi(t)dt$	$\phi(t)$	$\phi^{(2)}(t)$	$\phi^{(3)}(t)$	$\phi^{(4)}(t)$	t
3.00	.498650	.00443	.03545	− .07977	.13296	3.00
3.01	.498694	.00430	.03466	− .07845	.13214	3.01
3.02	.498736	.00417	.03389	− .07713	.13128	3.02
3.03	.498777	.00405	.03312	− .07582	.13038	3.03
3.04	.498817	.00393	.03237	− .07452	.12944	3.04
3.05	.498856	.00381	.03163	− .07323	.12847	3.05
3.06	.498893	.00370	.03090	− .07195	.12747	3.06
3.07	.498930	.00358	.03019	− .07068	.12643	3.07
3.08	.498965	.00348	.02949	− .06943	.12536	3.08
3.09	.498999	.00337	.02880	− .06818	.12426	3.09
3.10	.499032	.00327	.02813	− .06694	.12313	3.10
3.11	.499065	.00317	.02746	− .06571	.12198	3.11
3.12	.499096	.00307	.02681	− .06450	.12080	3.12
3.13	.499126	.00298	.02617	− .06330	.11960	3.13
3.14	.499155	.00288	.02555	− .06211	.11838	3.14
3.15	.499184	.00279	.02493	− .06093	.11714	3.15
3.16	.499211	.00271	.02433	− .05977	.11588	3.16
3.17	.499238	.00262	.02374	− .05861	.11460	3.17
3.18	.499264	.00254	.02316	− .05747	.11330	3.18
3.19	.499289	.00246	.02259	− .05635	.11199	3.19
3.20	.499313	.00238	.02203	− .05523	.11066	3.20
3.21	.499336	.00231	.02148	− .05413	.10933	3.21
3.22	.499359	.00224	.02095	− .05305	.10798	3.22
3.23	.499381	.00216	.02042	− .05198	.10662	3.23
3.24	.499402	.00210	.01991	− .05092	.10525	3.24
3.25	.499423	.00203	.01940	− .04987	.10387	3.25
3.26	.499443	.00196	.01891	− .04884	.10249	3.26
3.27	.499462	.00190	.01843	− .04782	.10110	3.27
3.28	.499481	.00184	.01795	− .04682	.09970	3.28
3.29	.499499	.00178	.01749	− .04583	.09830	3.29
3.30	.499517	.00172	.01704	− .04485	.09690	3.30
3.31	.499534	.00167	.01659	− .04389	.09549	3.31
3.32	.499550	.00161	.01616	− .04294	.09409	3.32
3.33	.499566	.00156	.01573	− .04201	.09268	3.33
3.34	.499581	.00151	.01532	− .04109	.09128	3.34
3.35	.499596	.00146	.01491	− .04018	.08987	3.35
3.36	.499610	.00141	.01451	− .03929	.08847	3.36
3.37	.499624	.00136	.01413	− .03841	.08707	3.37
3.38	.499638	.00132	.01375	− .03755	.08567	3.38
3.39	.499650	.00127	.01338	− .03670	.08428	3.39
3.40	.499663	.00123	.01301	− .03586	.08290	3.40
3.41	.499675	.00119	.01266	− .03504	.08151	3.41
3.42	.499687	.00115	.01231	− .03423	.08014	3.42
3.43	.499698	.00111	.01197	− .03344	.07877	3.43
3.44	.499709	.00107	.01164	− .03266	.07741	3.44
3.45	.499720	.00104	.01132	− .03189	.07606	3.45
3.46	.499730	.00100	.01100	− .03114	.07471	3.46
3.47	.499740	.00097	.01070	− .03040	.07338	3.47
3.48	.499749	.00094	.01040	− .02967	.07205	3.48
3.49	.499758	.00090	.01010	− .02895	.07074	3.49

Areas under the Curve $y = \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, the Function $\phi(t)$, and its Second, Third, and Fourth Derivatives

t	$\int_0^t \phi(t)dt$	$\phi(t)$	$\phi^{(2)}(t)$	$\phi^{(3)}(t)$	$\phi^{(4)}(t)$	t
3.50	.499767	.00087	.00982	− .02825	.06943	3.50
3.51	.499776	.00084	.00954	− .02757	.06814	3.51
3.52	.499784	.00081	.00927	− .02689	.06685	3.52
3.53	.499792	.00079	.00900	− .02623	.06558	3.53
3.54	.499800	.00076	.00874	− .02558	.06432	3.54
3.55	.499807	.00073	.00849	− .02494	.06308	3.55
3.56	.499815	.00071	.00824	− .02432	.06184	3.56
3.57	.499822	.00068	.00800	− .02370	.06062	3.57
3.58	.499828	.00066	.00777	− .02310	.05941	3.58
3.59	.499835	.00063	.00754	− .02252	.05821	3.59
3.60	.499841	.00061	.00732	− .02194	.05703	3.60
3.61	.499847	.00059	.00710	− .02138	.05586	3.61
3.62	.499853	.00057	.00689	− .02082	.05471	3.62
3.63	.499858	.00055	.00669	− .02028	.05357	3.63
3.64	.499864	.00053	.00649	− .01975	.05244	3.64
3.65	.499869	.00051	.00629	− .01923	.05133	3.65
3.66	.499874	.00049	.00610	− .01873	.05023	3.66
3.67	.499879	.00047	.00592	− .01823	.04915	3.67
3.68	.499883	.00046	.00574	− .01774	.04808	3.68
3.69	.499888	.00044	.00556	− .01727	.04703	3.69
3.70	.499892	.00042	.00539	− .01680	.04599	3.70
3.71	.499896	.00041	.00522	− .01635	.04497	3.71
3.72	.499900	.00039	.00506	− .01590	.04396	3.72
3.73	.499904	.00038	.00491	− .01547	.04297	3.73
3.74	.499908	.00037	.00475	− .01504	.04200	3.74
3.75	.499912	.00035	.00461	− .01463	.04103	3.75
3.76	.499915	.00034	.00446	− .01422	.04009	3.76
3.77	.499918	.00033	.00432	− .01383	.03916	3.77
3.78	.499922	.00031	.00419	− .01344	.03824	3.78
3.79	.499925	.00030	.00405	− .01306	.03734	3.79
3.80	.499928	.00029	.00392	− .01269	.03646	3.80
3.81	.499930	.00028	.00380	− .01233	.03559	3.81
3.82	.499933	.00027	.00368	− .01198	.03473	3.82
3.83	.499936	.00026	.00356	− .01164	.03389	3.83
3.84	.499938	.00025	.00344	− .01130	.03307	3.84
3.85	.499941	.00024	.00333	− .01098	.03226	3.85
3.86	.499943	.00023	.00322	− .01066	.03146	3.86
3.87	.499946	.00022	.00312	− .01035	.03068	3.87
3.88	.499948	.00021	.00302	− .01004	.02991	3.88
3.89	.499950	.00021	.00292	− .00975	.02916	3.89
3.90	.499952	.00020	.00282	− .00946	.02842	3.90
3.91	.499954	.00019	.00273	− .00918	.02770	3.91
3.92	.499956	.00018	.00264	− .00891	.02699	3.92
3.93	.499958	.00018	.00255	− .00864	.02630	3.93
3.94	.499959	.00017	.00247	− .00838	.02562	3.94
3.95	.499961	.00016	.00238	− .00813	.02495	3.95
3.96	.499962	.00016	.00230	− .00788	.02430	3.96
3.97	.499964	.00015	.00223	− .00764	.02366	3.97
3.98	.499966	.00014	.00215	− .00741	.02303	3.98
3.99	.499967	.00014	.00208	− .00718	.02242	3.99

INDEX

Rietz, H. L. (Ed.)
Handbook of mathematical
statistics.